COMPARISON OF GPS-EQUIPPED VEHICLES AND ITS ARCHIVED DATA FOR THE ESTIMATION OF FREEWAY SPEEDS

A Thesis
Presented to
The Academic Faculty

By

Jaesup Lee

In Partial Fulfillment
Of the Requirement for the Degree
Master of Science in Civil and Environmental Engineering

Georgia Institute of Technology

May 2007

COMPARISON OF GPS-EQUIPPED VEHICLES AND ITS ARCHIVED DATA FOR THE ESTIMATION OF FREEWAY SPEEDS

Approved by:

Dr. Randall Guensler, Advisor
School of Civil and Environmental Engineering
Georgia Institute of Technology

Dr. Michael Hunter
School of Civil and Environmental Engineering
Georgia Institute of Technology

Dr. Michael Rodgers
School of Civil and Environmental Engineering
Georgia Institute of Technology

Date Approved: April 9, 2007

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Video image detection system (VDS) equipment provides real-time traffic data for monitored highways directly to the traffic management center (TMC) of the Georgia Department of Transportation (GDOT). However, at any given time, approximately 30 to 35% of the 1,600 camera stations (STNs) fail to work properly. The main reasons for malfunctions in the VDS system include long term road construction activity (eliminating stations from service or severing communications with the stations, and technical) and operational limitations. Thus, providing alternative data sources for offline VDS stations and developing tools that can help detect problems with VDS stations can facilitate the successful operation of the TMC.

To estimate the travel speed of non-working STNs, this research examined global positioning system (GPS) data from vehicles using the ATMS-monitored freeway system as a potential alternative measure to VDS. The goal of this study is to compare VDS speed data for the estimation of the travel speed on freeways with GPS-equipped vehicle trip data, and to assess the differences between these measurements as a potential function of traffic and roadway conditions, environmental, conditions, and driver/vehicle characteristics. A preliminary analysis shows that the mean of GPS speeds is higher than that of the VDS, and the standard deviation of the GPS speed is equal to or lowers than the VDS. The difference between GPS and VDS speeds is affected by various factors such as congestion level (expressed as level of service), onroad truck percentage,

facility design (number of lanes and freeway sub-type), posted speed limit, weather, daylight, and time of day. The relationship between monitored speed difference and congestion level was particularly large and was observed to interact with most other factors.

This study utilized descriptive statistics followed by classification and regression tree (CART) analysis to identify significant variables and assess interaction effects with respect to the noted difference between VDS and GPS speed. CART analysis results indicated that driver age was the most relevant variable in explaining variation for the southbound of freeway dataset and freeway sub-type, speed limit, driver age, and number of lane were the most influential variables for the northbound of freeway dataset. The combination of several variables had significant contribution in the reduction of the deviation for both the northbound and the southbound dataset. Although this study identifies potential relationships between speed difference and various factors, the results of the CART analysis should be considered with the driver sample size to yield statistically significant results. Expanded sampling with larger number of drivers would enrich this study results.

CHAPTER ONE

INTRODUCTION

1.1 Background

According to the Texas Transportation Institute (TTI) and Cambridge Systematics Inc. (Texas Transportation Institute and Cambridge Systematics Inc., 2005), congestion levels in major cities around the United States are getting worse. Because high congestion on one highway segment can easily transfer to other highways, and because high congestion levels can also cause the likelihood increase of traffic incidents, traffic monitoring systems are more important than even before (Texas Transportation Institute and Cambridge Systematics Inc., 2005). The acquisition of reliable traffic data is the critical factor for evaluating traffic performance. Thus, various traffic performance measures such as travel time, speed, and delay are crucial input data to the traffic management system.

Previous research has been conducted in the areas of travel time studies estimation and prediction, traffic incident detection, real-time traffic information dissemination, and in-vehicle route guidance. For example, one means of reducing traffic congestion is to provide more accurate and reliable traffic information to drivers via traffic data collection technologies. Older inductive loop detection systems that use a wire coil and magnetic field for vehicle detection are prevalent in most urban areas.

Plus, many start-of-art technologies that use video processing are deployed in traffic detection system. Video systems and other non-intrusive technologies for traffic detection have been widely used (Minnesota DOT and SRF Consulting Group, 2002).

Transportation management centers (TMCs), or traffic management centers, operated by state departments of transportation (DOT) contractors have performed a critical role for successful implementation of congestion monitoring system. Most TMCs are monitoring the traffic conditions on freeways and on some major arterials and gathering real-time information from many traffic-related sources such as video monitoring and detection systems, freeway call boxes, 911 calls, officers on patrol, and motorist cellular calls along the freeways and major arterials (Georgia State DOT, 2007). In particular, various non-intrusive technologies such as passive infrared, active infrared, magnetic, radar, Doppler microwave, pulse ultrasonic, passive acoustic, and video image processing were recently applied for detecting traffic condition (Mimbela and Klein, 2000; Minnesota DOT and SRF Consulting Group, 2002; Oregon DOT, 2005).

The freeway monitoring system operated by the Georgia DOT and known as "The Georgia NaviGAtor," employs a widely-deployed video detection system (VDS). Surveillance data from VDS cameras are the primary source for generating traffic performance measures such as travel time, speed, and delay in Georgia NaviGAtor. However, many stations in the Georgia Navigator system do not work properly all of the time. Figure 1-1 shows that the malfunctioning ratio of the Georgia Navigator system were 27% during the am peak of September 1st and 35% during the pm peak of October

3rd, 2006. The main reasons for malfunctions in the VDS system are long term road construction projects and technical and operational limitations. Thus, the estimation of non-working STNs is significant for the successful operation of the TMC.

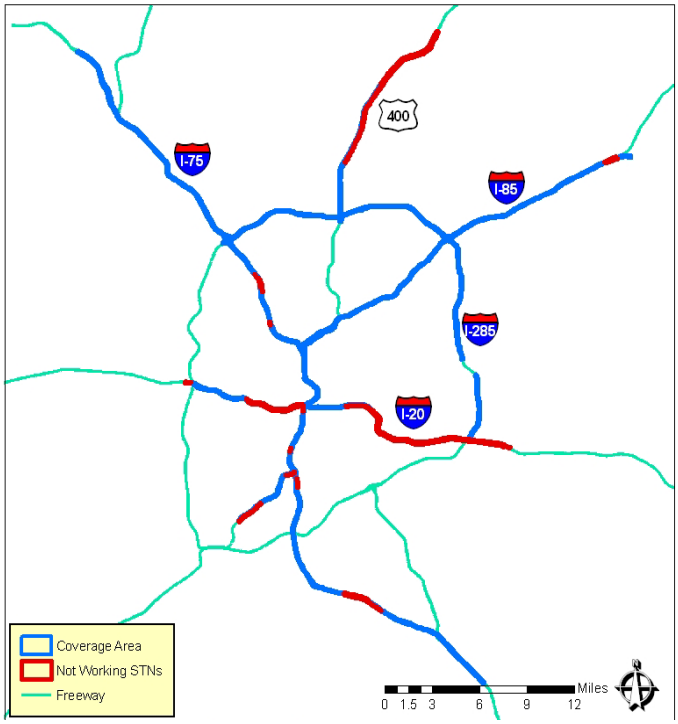


FIGURE 1-1 Non-Working Status on Georgia NaviGAtor

Therefore, reliable traffic data acquisition is still a major concern to transportation system operators and decision makers. Even though advanced technologies are deployed in ITS applications, TMCs still have technical and operational issues to consider regarding traffic data quality.

1.2 Research Objectives

To estimate the travel speed of non-working STNs on freeways, this research examined global positioning system (GPS) data as an alternative measure to the VDS. The goal of this study is to evaluate the differences between VDS speed data and GPS data for the estimation of the travel speed as a function of traffic conditions, roadway design parameters, environmental conditions, and driver/vehicle characteristics.

To achieve this goal, this study investigates the difference between GPS and VDS speed associated with various factors affecting speed differences. The major research objectives of this research effort are as follows:

- Investigate the general relationships between VDS and GPS speed
- Investigate potential relationships between the speed difference (GPS and VDS) and traffic conditions
- Investigate potential relationships between the speed difference (GPS and VDS) and roadway characteristics
- Investigate potential relationships between the speed difference (GPS and VDS) and environmental characteristics
- Investigate potential relationships between the speed difference (GPS and VDS) and driver/vehicle characteristics
- Investigate potential relationships between the speed difference (GPS and VDS) and potential explanatory variables in the above categories with interaction effects

1.3 Research Methods

This research utilizes the subset of GPS data including driver/vehicle information obtained from the instrumented vehicles traveled the metro Atlanta region. To compare with GPS speed and define traffic conditions, VDS data from Georgia department of transportation (GDOT) transportation management center (TMC) are also utilized. In addition to the GPS and VDS data, roadway characteristics data including GDOT roadway characteristics (RCLINK) database and aerial photos from US Geological Survey (USGS), environmental characteristics data including precipitation and sunrise/sunset data are also utilized in order to analyze the effect on the difference between GPS and VDS speed. This research utilizes the Kalmogorov-Smirnov (KS) test to examine the magnitude of differences among the distributions. This research also utilizes classification and regression tree (CART) analysis to analyze the effect of combined variables.

1.4 Thesis Outline

Following the introductory chapter, Chapter 2 reviews the traffic data collection technologies used in the field (video detection systems detail, GPS technologies for traffic studies, and speed estimation studies). Chapter 3 discusses the data and preparation process employed in this study. Chapter 4 evaluates the quality of GPS data through the reduction procedure with various considerations. Chapter 5 analyzes the independent effects of nine variables on the difference between GPS and VDS speed. Chapter 6 analyzes the combined effects of nine variables on the difference between GPS

and VDS speed. Finally, Chapter 7 summarizes the findings from this research effort and provides further research suggestions.

CHAPTER TWO

LITERATURE REVIEW

2.1 Data Collection Technologies

A variety of traffic detection technologies have been applied with various degrees of success in different fields for detecting vehicles and determining volume, occupancy, and speed. Data collection technologies can be divided into three categories: point detection, beacon-based probe vehicle technology, and non-traditional probe vehicle performance (Texas Transportation Institute and Cambridge Systematics Inc., 2005).

2.1.1 Point Detection

Point detection techniques place surveillance equipment at a specific location and measure traffic performance including speed, volume, and occupancy. Loop detectors, video image detection systems, and microwave radar systems are the most common equipment. Measurements taken at consecutive locations are used to measure travel time along each specific segment. Intrusive detectors such as inductive loops have been widely used for decades (Martin and Feng, 2003). When a vehicle passes over a loop or pauses in an inductive loop area, loop inductance is reduced and oscillator frequency is increased. Thus vehicle's presence is determined when frequency change exceeds the

threshold set by the sensitivity setting. However, an inductive loop detector has significant disadvantages compared to other point detection systems. First, a loop detector is expensive to purchase. Second, a loop detector requires lane closure for installation and maintenance. Third, loop wires can break during freeze/thaw climate conditions, poor pavement conditions, and if improperly installed. Because video cameras are installed above ground or on the roadside, no traffic closure is required for installation and maintenance of the system (Martin and Stevanovic, 2004). Microwave radar technology was used to detect objects during the World War II. Doppler microwave detectors, a popular type of microwave detector, transmit low energy microwave radiation to the detection zone. Using the Doppler effect, motion of a vehicle causes a frequency shift and detectors measure this shift to determine the passage and speed of vehicle (Martin and Feng, 2003). Video detection systems were introduced due to the limitations of inductive loop detection technologies and microwave systems and will be discussed in detail later in this chapter.

2.1.2 Beacon-based Probe Vehicle Technology

Beacon-based probe vehicle technology is generally used in electronic toll collection systems. A beacon system interrogates electronic vehicle tags when a vehicle passes the reader location, and also records the particular point and time on the roadway for the automated billing of that vehicle. Thus, since travel time and delay between specific two points can be obtained, travel time information is more accurate than data estimated from point detectors. However, beacon-based detectors do not provide traffic information within monitoring segments such as vehicle trace information and delay.

Another major limitation of this method is that most tag-based data collection systems can only identify a sample of all vehicles, thus traffic volumes must be collected from other sources.

2.1.3 Non-traditional Probe Vehicle Performance

Non-traditional probe vehicle performance systems provide travel time, speed and delay information without the installation of the infrastructure for point-based or beacon-based detection systems throughout networks. Two common technologies for non-traditional prove vehicle performance systems are cell phone tracking and GPS-equipped vehicle technologies. The cell phone tracking technology determines the travel speed by tracking the approximate location of cell phones. Cell phone tracking technology can provide huge advantages compared to other technologies given that the presence of cell phones in vehicles is pervasive. However, the accuracy of these data for use in determining speed profiles is still being studied. The research for cell phone tracking system assumed that the technology of cell phone positioning systems has a very high order of accuracy but have an error of the order of 100 meters RMS. The ambiguous phone positioning systems cause errors such as the assignment of a vehicle to the wrong road, and mistakes in the direction of travel (Schneider and Mrakotsky, 2005; Ygnace and Drane, 2001)

Another technology is GPS-equipped vehicles with wireless data transmission technology as used in this study. GPS equipment installed inside of a vehicle collects and reports vehicle operation information such as location (latitude and longitude),

heading, and speed while vehicle engine turned on. The major challenges for this technology is that a large number of vehicles should be equipped with a GPS device to obtain an unbiased and reliable roadway performance measures considering the temporal and spatial diversity of vehicle use. Another disadvantage to probe vehicle performance technology is that, like toll tag tracking, each probe vehicle can not provide information about the level of roadway condition such as vehicle volume or density. Thus, supplemental traffic volume data must be obtained when probe vehicles are the primary source of performance measures.

2.2 Video Image Detection System

2.2.1 Specification of Technologies

The video detection system combines real-time image processing with computerized pattern recognition technologies (Martin and Stevanovic, 2004). Generally, four major factors should be considered for the desirable video detection system: camera installation, camera height, field of view calibration, and adjusting the focus because they are closely related to the data accuracy generated by the VDS system.

Cameras installed along the roadside capture real-time video images of traffic data. Then captured images are sent to the TMC via fiber optic cable and recorded on video home system (VHS) video tapes. The video images from these cameras are processed and analyzed with image processing units. The VDS estimates traffic volume, speed, and density by detecting pixel color or shade changes over user-defined detection regions. After signal analysis is completed, estimated traffic information are recorded

and distributed with the communication between other controllers. Estimated traffic information are aggregated into various time intervals such as 20 seconds and one minute, three minutes, 15 minutes, and one hour (Grant et al., 1999; Martin and Stevanovic, 2004). Figure 2-1 shows the whole procedure of video detection system from cameras in the field to the internet for the public dissemination of the data.

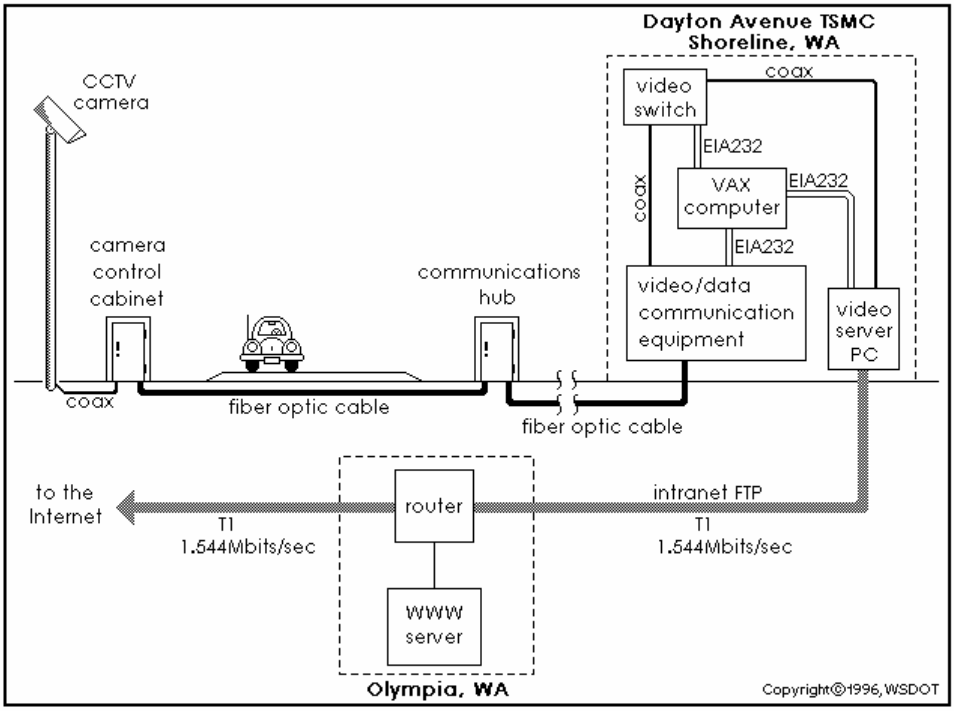


FIGURE 2-1 Flow Chart VDS System Working from Field to the Internet (Washington State DOT, 2007)

Like the VDS system of the Washington DOT, Georgia NaviGAtor collects the majority of traffic information using the video image detection system. The Georgia TMC currently employs two types of cameras for VDS: the AutoScope and the Traficon video processors (URS Corporation and GeoStats Inc., 2003). There are 341 CCTV

cameras on Atlanta freeways and 207 CCTV cameras on arterials to monitor traffic conditions. Each camera has full color and pan/tilt/zoom (PTZ) capability, VDS camera installed approximately one third mile at mainlines, HOV lanes, ramps, and interchanges on the freeway, detect traffic flows via image processing. The VDS covers 220 miles of freeway with more than 1,600 locations in Atlanta and Macon.

As shown in Figure 2-2, Georgia NaviGAtor provides real-time traffic information in 3-minute interval including congestion levels, travel times, accidents, lane blockages, weather, and construction schedules. Traffic information is made available via the internet (http://data.georgia-navigator.com), mobile internet (PDAs), web-enabled mobile phones, and the variable message sign (VMS) along the freeways. TMC camera stations monitor data within into several sub-areas with travel speed data in each sub-area updated approximately every 3-minutes (URS Corporation and GeoStats Inc., 2003).

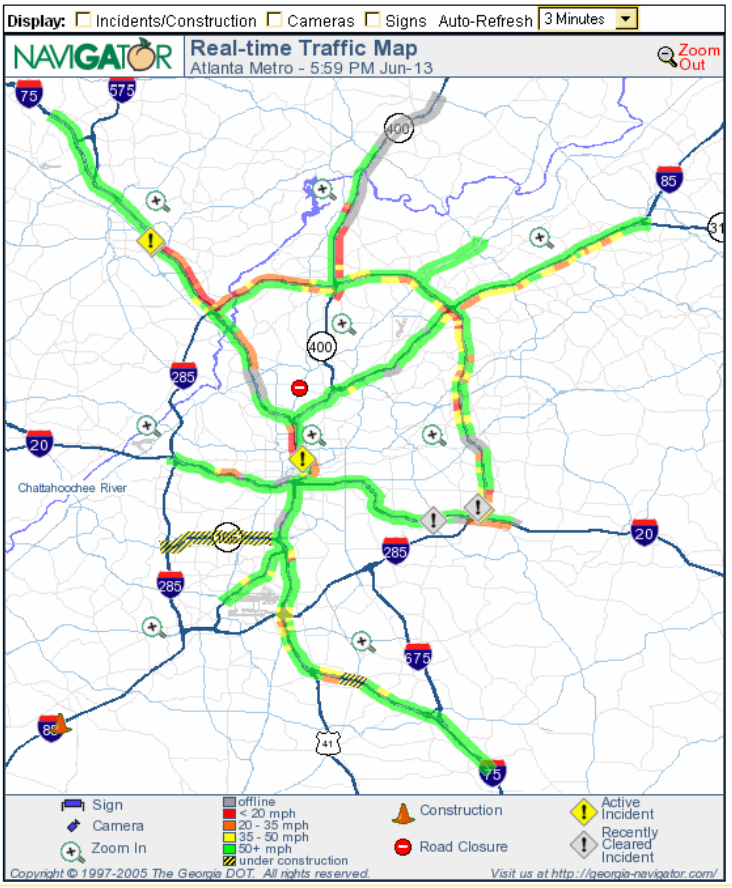


FIGURE 2-2 Georgia NaviGAtor System on the Website

2.2.2 Factors affecting VDS error

Most TMCs are currently performing research and operational tests, preparing technical guidance and recommended practices, developing training, or pursuing technology transfer initiatives (TMC Pooled-Fund Study, 2007). Previous research evaluated the traffic data quality of numerous detection systems under various conditions including weather, road geometric, traffic level, mounting configurations, non-motorized traffic detection, and train detection (Grant et al., 1999; Martin and Stevanovic, 2004;

Mimbela and Klein, 2000; Minnesota DOT and SRF Consulting Group, 2002). Martin and Stevanovic (2004) categorized the sources of errors in video detection into three as shown in Figure 2-3: camera installation error, detector file creation error, and algorithm error in vision processor.

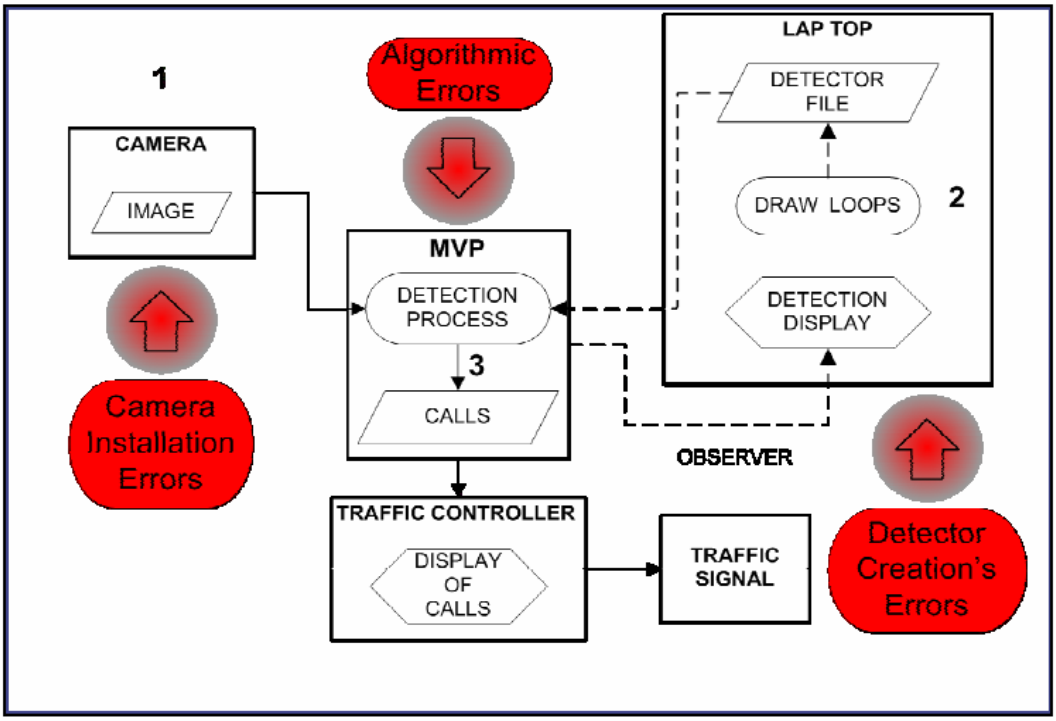


FIGURE 2-3 Source of Errors in the Video Detection System (Martin and Stevanovic, 2004)

Martin and Stevanovic (2004) found that the accuracy of VDS is dependent upon camera installation factors such as the camera height, location, and angle above roadway and environmental factors such as rain, sun intensity, and day/night also affect vehicle detection accuracy. Video systems tend to work poorly in low visibility weather conditions such as heavy snow or thick fog. In addition, video detection system requires

maintenance in short period of time such as six months because dirt and water can affect the image quality causing detection accuracy.

Middleton and Parker (2004) evaluated the accuracy of video detection systems under various conditions and found that lighting, number of lanes, congestion, pavement structure, location, geometry, and traffic mix affect the accuracy of video detection system data. They recommend that the ideal site for the detector installation is the location that never experiences stop-and-go traffic condition, which means traffic congestion is very critical to VDS data accuracy. With respect to VDS accuracy, Washington State DOT (2007) stated:

> *"The accuracy of calculated travel times varies depending on congestion. The equipment used to estimate speeds on the freeway becomes inaccurate when traffic exceeds 60 mph or drops into stop-and-go congestion"*

Some research evaluating the accuracy of video detection data of the stop bar at signalized intersections, found that the critical element of the VDS application at signalized intersection is occlusion. For example, during a red signal, heavy truck in the lane being counted blocks small vehicles in an adjacent lane such as auto vehicle for VDS system to detect. (Rhodes et al., 2005; Tian, 2006). Similar situation can occur on freeways especially congested time period such as stop-and-go traffic condition.

Grant, et al. (1999) found that camera image motion, camera angle, slanted camera view, poor lighting conditions, heavy traffic volumes, inclement weather, media

used in collecting data, the placement of detectors in a particular lane also affects count accuracy. They also found that camera angle affects traffic counts in adjacent lanes due to the occlusion. For example, on a six-lane segment, the lane closest on the camera only varied 4.8%, while the other lanes had discrepancies in truck counts of 55% to 84%. In addition, more than 40% of the locations underestimated VDS average speeds compared to ground truth spot speeds. The estimated average travel speeds from the VDS were significantly different with measured speeds on the freeways. Thus, they did not recommend use VDS data for vehicle classification of trucks due to high error level.

2.3 GPS Technologies in Traffic Studies

Among the various travel data collection methods, the GPS technology has been the most common choice in transportation research because the systems provide additional useful data, such as start and stop points of a trip, travel routes, travel time, second-by-second speed, acceleration rates, etc. The major advantages of the GPS technology are that GPS equipment is very flexible to deploy and use and GPS equipment requires low capital and installation cost. From the installation, operation, and maintenance cost perspectives, GPS technology is a significantly competitive alternative when compared with existing traffic detection systems. The minimal labor and hardware requirement for the transportation data collection and analysis are also significant advantages compared to traditional methods (Zito and Taylor, 1994).

Recently, as part of an effort in obtaining traffic data, GPS-equipped probe vehicle collect traffic data in real time and integrated with geographic information system

(GIS) application. The integration of the GIS technology allows the effective display of collected GPS data and provides more powerful analysis results. More specifically, travel time, speed, and delay analysis using an integrated GIS/GPS system have recently become available for use in conducting different transportation studies. Faghri and Hamad (2002) performed a comparative statistical analysis with data collected by GPS method and by traditional methods. They argue that GPS data were at least as accurate as that of conventional methods, and GPS data collection method was 50% more efficient in terms of manpower (Faghri and Hamad, 2002). Quiroga (1997) analyzed segment length with GPS-equipped data in a travel time study and found that relatively short segments (0.2-0.5 miles long) are needed to detect localized traffic effects.

Previous research applied GPS technologies in various studies:

- To measure the impact of construction activities (Garcia et al., 2006)
- To measure travel time delay along a freeway segment (Wang, 2004)
- To identify travel behavior information combined with driver information and trip purpose (Greaves and Somers, 2003; Wolf et al., 2004)
- To estimate macroscopic and microscopic traffic parameters in simulation models

Although current GPS devices have high accuracy, data accuracy and reliability of GPS data including systematic error and random error are still the main concerns to GPS data users (Ogle, 2005). While systematic error can be identified and corrected easily by applying screen out rules, random errors are more difficult to detect. Thus, statistical smoothing techniques may be applied to identify random error from huge raw data. Applying smoothing techniques is not only to minimize the impact of random

errors on the results but also to require much less processing time for detecting random errors than visual inspection (Jun, 2006).

2.4 Speed Estimation

2.4.1 Factors Affecting Driver's Speed Choice

The vehicle speed on the road depends on various factors related to the traffic conditions, environmental characteristics and driver/vehicle characteristics. Previous researchers have investigated a wide variety of factors that affect travel speed of drivers and speed distributions associated with the speed-flow relationship (Brilon and Ponzlet, 1996), traffic safety (Liang et al., 1998; Ogle, 2005), traffic operation and roadway design (Hoogendoorn, 2005; Hostovsky et al., 2004; Kanellaidis, 1995; Kyte et al., 2001), and vehicle emissions (Hallmark et al., 2004). They found that vehicle speed is affected by the following factors:

- Roadway characteristics such as speed limit and number of lanes
- Geometry of the roadway such as horizontal and vertical alignments for sight distance and roadway surface condition
- Traffic condition such as density of the traffic stream, traffic volume, and traffic mix such as heavy-duty truck percentage
- Weather and environmental factors reducing driver's visibility such as darkness, rain, snow, fog, dust, and wind.

Hoogendoorn (2005) investigated the free flow speed distribution in terms of vehicle type and travel lane-by-lane on motorway. He found that the travel speed of

personal vehicles is higher than trucks, and the speed of personal vehicles driving in the left lane is higher than person cars driving in the right lane. He also found that the speed of trucks in the left lane higher than trucks in the left lane and the speeds of trucks on either lane are much lower than the personal vehicles. Hallmark, Knapp et al. (2004) compared the spot speeds of passenger cars, light-duty trucks, SUVs, and vans in order to separately estimate the emission rates for each vehicle class. However, they found that the differences in mean speeds within light-duty vehicle classes were not statistically significant.

CHAPTER THREE

DATA COLLECTION

3.1 Study Area

The study area GS-400 corridor consists of a 12-mile length of freeway segments located on the north of the I-285 freeway circle encompassing the Atlanta metropolitan area (Figure 3-1). This corridor has 23 on/off-ramps in each direction and does not contain HOV lanes. The posted speed limits on the corridor are 55 mph and 65 mph. The number of lanes is four at the southern part of the corridor, near the interchange with the I-285 freeway, and decreases as two at the northern part of study corridor.

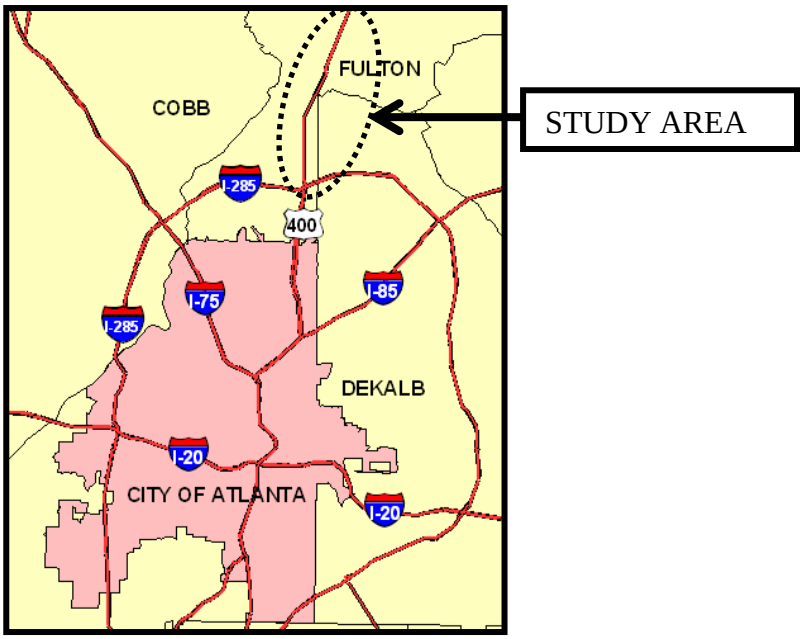


FIGURE 3-1 Study Area GA-400

3.2 Data Preparation

This study collected data of GPS and VDS speed for eight months from January to August 2004. For comparisons between GPS and VDS speed, five different datasets were obtained from the Georgia DOT, Commute Atlanta Program (http://commuteatlanta.ce.gatech.edu), USGS, and NOAA.

- VDS traffic data from TMC, Georgia DOT
- GPS vehicle trip data from Commute Atlanta Program
- GPS driver/vehicle information from Commute Atlanta Program
- Roadway Characteristics
  - One-foot resolution aerial photo from U.S. Geological Survey (USGS)
  - Roadway Characteristics Table (RCLINK) from Georgia DOT
- Environmental characteristics
  - Weather and daylight saving data from national oceanic and atmospheric administration (NOAA)

3.2.1 VDS Traffic Data

The majority of the VDS systems used by Georgia NaviGAtor consist of count station data acquisition (CSDA) systems that collect and archive video-detected traffic data using the AutoScope technology. Georgia DOT operates transportation sensor system (TSS) units as a replacement to CSDA system and installed Traficon VIDS unit along several sections including the GA400, I-75 south of I-285, and I-285 NE (URS Corporation and GeoStats Inc., 2003). To reduce the storage space required for raw

traffic data, the TMC archives raw data in three-minute, 15-minute, and one-hour interval. Operating under an agreement with the Georgia DOT, the School of Civil and Environmental Engineering at the Georgia Institute of Technology (Georgia Tech) receives the TSS archived data at 20-second intervals. This study utilizes 84,326,400 STN-based travel speed records from January to August 2004.

Along this corridor, 77 VDS cameras stations (39 for northbound and 38 for southbound) are installed approximately every 0.3 miles. Since VDS cameras measure traffic conditions on every 0.3 mile segment, traffic conditions within each segment are assumed to be the same. The TMC generates ITS archived data every 20 seconds for all STNs, which assumes that traffic characteristics of all vehicles traveled within the same STN are constant over these 20 seconds. In other words, all VDS stations have one speed value for each single 20-second time interval. CCTV cameras in the field capture the traffic condition and perform video image processing in real-time. Then, processed detection data are transferred to the TMC server. Processed real time data were transferred into the Georgia Tech server at the same time. This study utilizes the archived data of 2004, which were already stored in Georgia Tech server.

After obtaining compressed trip files from Georgia Tech Transportation Group sever, traffic information data were extracted from the compressed trip files for each STN, including STN ID, date, time (hhmmss), speed, density, and volume counts for autos, vans, trucks etc. By using extracted traffic information, matrices for the travel speed and density was developed from January 1st to August 31st at 20-second intervals. Only

valid detector data were used in this process. The data quality control procedure for the raw data was already applied by the TMC. The TMC applied data screening rules for rejecting 20 seconds data samples from the system as shown in Table 4-1. Any one of below cases was defined as abnormal data and removed from the original dataset.

TABLE 3-1 Data Screening Rules for Rejecting 20-second Data Samples

Case	Volume (veh/h)	Speed (mph)	Occupancy (%)
1	zero	zero	Zero
2	any	low (< 20)	medium (>=10 & < 30)
3	high (>= 2,700)	any	low or high (<10, > 50)
4	nearly zero (< 360)	nearly zero (<=10)	not nearly zero
5	too high (> 3600)	any	Any
6	any	any	too high (> 100)
7	any	too high (> 100)	Any

Source: (URS Corporation and GeoStats Inc., 2003)

3.2.2 GPS-equipped Vehicle Data

GPS-equipped vehicle trip data were obtained from the commuter choice and value pricing insurance incentive program (Commute Atlanta study), funded by Federal Highway Administration (FHWA) and the Georgia DOT implemented in 2003. The focus of Commute Atlanta program is to assess the effects how fixed vehicle operating costs converted into mileage-based and congestion-based operating costs. The researchers of the Commute Atlanta study installed GPS equipment in the light-duty vehicles of resident commuters through out the Atlanta metropolitan area in order to collect travel data. Figure 3-2 shows the distribution of participant drivers, which are

465 instrumented vehicles from 261 households within metro Atlanta 13 county area in 2004. The Commute Atlanta study is currently conducting variety of projects including research in the areas of travel behavior, value pricing, safety, traffic operations, and motor vehicle emissions. The driver information including age, gender, vehicle type, and vehicle year of each participant were combined with each driver's trip data and provide very useful insight in driver behavior analysis

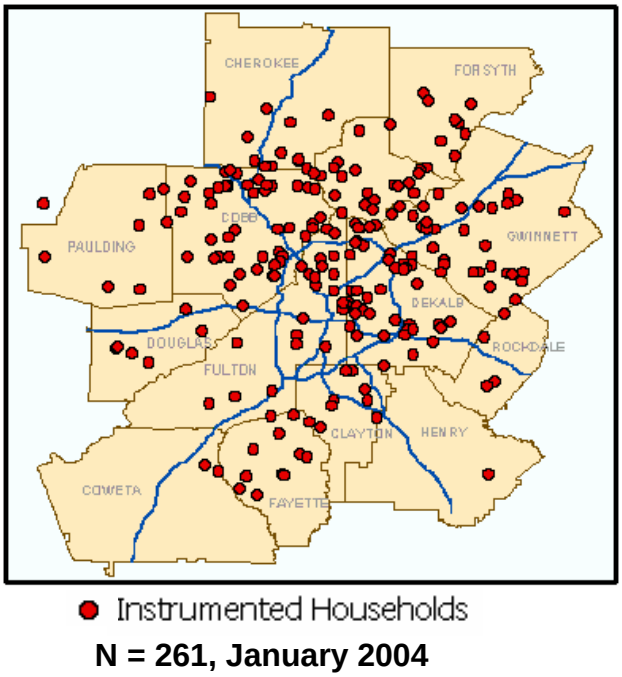


FIGURE 3-2 Location of the Households for Participating Commute Atlanta Study

To collect driver's trip data, the DRIVE Atlanta Laboratory at Georgia Tech developed a wireless data collection system known as the GT Trip Data Collector (GT-TDC). With the ignition of the engine, trip data collector starts to run and continuously records engine information and vehicle's movement on a second-by-second basis until the engine is turned off. The GT-TDC collects second-by-second vehicle activity data

including vehicle position (latitude and longitude) and vehicle velocities. In addition, the GT-TDC collects ten engine-operating parameters from the onboard diagnostics (OBD) system in post-1996 model year vehicles and monitors vehicle speed at 4Hz from the vehicle speed sensor (VSS).

Collecting trip data including vehicle and engine parameters from GPS equipment, OBD, and VSS systems are integrated into the trip files and encrypted in memory space installed inside of the GPS equipment. Recorded trip data were transmitted to the central server system at Georgia Tech periodically (10 pm to 6 am during weekdays and the anytime of weekend) by using a wireless data transmit system via a cellular connection. Figure 3-3 illustrates the appearance of the GT-TDC and its accessories including GPS receiver, cellular transceiver, the onboard diagnostic connector, and VSS connector.

FIGURE 3-3 GT Trip Data Collector (Jun, 2006)

3.2.3 GPS Driver and Vehicle Data

The Commute Atlanta study provides all GPS trip data for the GA 400 corridor that are used in this study. 138,284 STN-based VDS speed measurements are compared to the corresponding GPS measurements collected from 209 vehicles in the Commute Atlanta Program between January and August 2004. The 209 vehicle-driver dataset consists of 120 Autos, 21 Vans, 38 SUVs, and 30 Pick-up Trucks. Figure 3-4 shows the distributions of driver ages by vehicle type and gender.

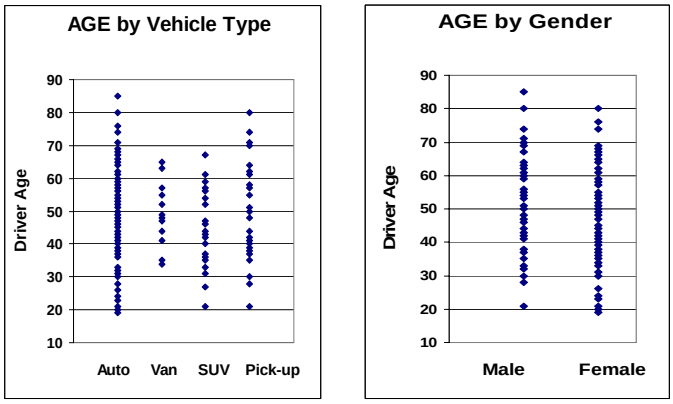


FIGURE 3-4 Driver Distributions by Vehicle Type and Gender

3.2.4 Roadway Characteristics

To parameterize roadway characteristics, this study utilizes roadway characteristics (RC) table and the aerial photos from the U.S. Geological Survey (USGS). The RC table contains the information of roadways including number of lanes, speed limit, and link length, etc. One-foot resolution aerial photos obtained from the USGS allow researchers to ensure that detailed roadway characteristics such as exact location of the gore at off-ramp and length of each the on/off-ramps are recorded. The number of

lanes and the speed limit data for the study corridor were obtained from the RC table. The study corridor consists of 2, 3, and 4 lanes segments with 55 and 65 mph speed limits as shown in Table 3-2.

TABLE 3-2 Number of STNs by Roadway Conditions

	Basic	On-Ramp	Off-ramp	Total
2lane - 65mph	5	3	2	10
3lane - 65mph	10	8	4	22
4lane - 65mph	21	2	4	27
4lane - 55mph	8	5	5	18
Total	44	17	15	77

Freeway sub-types are categorized as basic, on-ramp, and off-ramp segments based upon the existence of the on/off-ramp and their effective length (1,500 ft.) If the majority of data points within a STN overlap with on/off-ramp area, this STN categorized as on/off-ramp segment. The study corridor consists of 44 basic STNs, 17 on-ramp STNs, and off-ramp 15 STNs. In addition, 2-lane, 3-lane, and 4-lane of 65 mph speed limit STNs are 10, 22, and 27, respectively. The study corridor has 59 out of 77 STNS with 65 mph speed limit and 45 out of 77 STNs that are 4-lane segments.

3.2.5 Environmental Characteristics

Because poor weather conditions affect traffic flow, this study considers inclement weather information. Kyte, et al. (2001) demonstrated that pavement conditions, visibility, and wind speeds affect the free-flow speed on freeways, and thus

should be considered in capacity or LOS analyses. To consider bad weather conditions, this study utilized precipitation data from National Oceanic & Atmospheric Administration (NOAA). NOAA data shows that 69 days out of 244 from January to August have positive precipitation. Total number of inclement time is 481 hours out of 5,856 hours during the eight months. By using hourly based precipitation data, every single STN speed overlapped with inclement weather time period categorized as the data of inclement.

This study also utilized the information of sunrise and sunset time in order to identify the effect of a day and a night. Because sun glare in camera lenses during the sunrise and sunset period is significant and may affect the accuracy video detection system, the time periods of 30 minutes before sunset and 30 minutes after sunrise are also categorized as night time periods. As results, total 78,018 STN-based trips consist of 742,331 for fine day and 3,785 for inclement day. In addition, the dataset for this study consists of 65,869 trips for daytime and 12,149 for nighttime.

CHAPTER FOUR

DATA DEVELOPMENT AND QUALITY CONTROL

As discussed in Chapter three, although current GPS devices have high level of accuracy, GPS data contains systematic error and random error. While systematic error can be identified and corrected by applying screening rules, random errors are more difficult to detect. Thus, statistical smoothing techniques may be applied to identify random error from the huge raw dataset. This chapter describes the data reduction procedures used to obtain more reliable datasets for the speed analysis.

4.1 GPS Data Reduction Process

GPS data reduction was conducted in order to screen out erroneous GPS raw data. The whole procedure from map matching step to mean speed calculation step is shown as Figure 4-1. The first step is to collect the GPS trip data from individual vehicles traveled in the study corridor within study period from January to August 2004. In this step, a conservative approach was applied in order to consider vehicle sharing between family members. Vehicle trip data with more than 5% vehicle sharing ratio between two adults within the same household were eliminated from the dataset.

As the second step, map matching process was applied to select the applicable vehicle trips traveled on the roadway segments of interest. This study deployed a series

of proprietary GIS-based methods to link the GPS data to the specific roadway segments. After STN-based GPS raw data, by using trip information such as direction and second-by-second time stamps of each GPS data point in single trip, irrelevant GPS data points were identified and eliminated form the STN-based GPS raw data (usually occurring at ramp areas connected to arterial over-passes or under-passes.

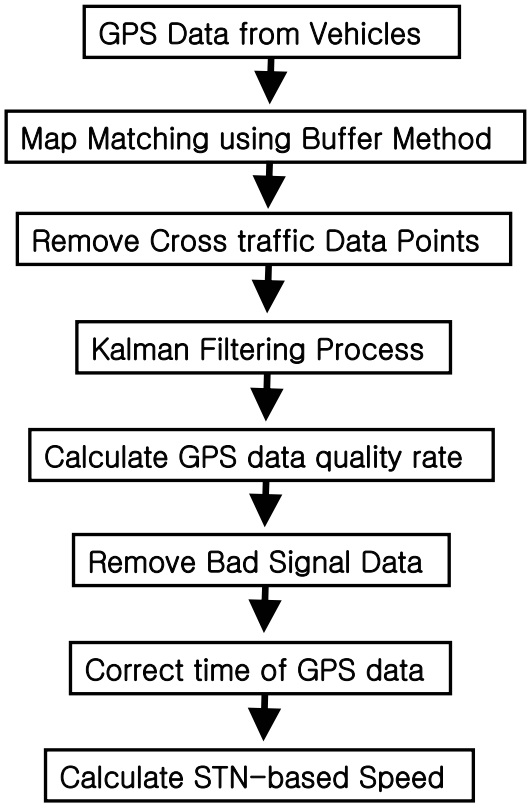


FIGURE 4-1 Whole procedure for GPS Data Reduction

Although the GPS technology provides the accurate speed data, the GPS is still contain various random errors (Ogle et al., 2002). Systematic errors in GPS data may be produced due to a low number of satellites and a relatively high position dilution of precision (PDOP) value, which is related to the satellite orientation on the horizon and other reasons that affect the precision and the accuracy of the GPS device. However, random error may be came from satellite orbits, a receiver, multi-path signal reflection etc. (Jun, 2006). Although systematic errors can be readily identified and removed, random errors are not easy to identify and correct as systematic errors. Thus, the application of statistical smoothing techniques is required in order to reduce the impact of random errors on the GPS data and improve the results of the study. Jun (2006) applied three methods in the smoothing process of GPS data including the least square spline approximation method, the Kernel-based smoothing method, and the Kalman filter method.

After comparison of accuracies between three methods, Jun (2006) recommended the Kalman filter method for smoothing GPS dataset. Thus, this study applied the Kalman filter method to reduce random errors from GPS dataset. The Kalman filter method consists of two recursive steps: prediction step and the correction step. In the prediction step, new value of next time ($t+1$) is estimated by the measurement of past time ($t-1$). Then, in correction step, estimated value is adjusted by the value of current time (t). By repeating recursive step, GPS data used in this study were refined.

After the Kalman filter smoothing process, the GPS data of bad quality were eliminated. Since the GPS data accuracy is usually determined by the number of satellite and the PDOP value, this study applied those two measurements as criteria for identifying bad data in this step. More specifically, all GPS data that do not satisfy two criteria (at least four satellites or PDOP value between 1 and 8) were defined as bad GPS data. Then, if valid GPS data point were less than 50%, the STN-based trip data were eliminated from the dataset.

To match GPS data with VDS data, the acquisition of accurate temporal and spatial information is critical. Because GPS raw data follows the UTC (Coordinated Universal Time known as Greenwich Mean Time) time, the GPS data of this study were adjusted to local time, a five hours difference. In addition to the UTC time correction, study data considered the daylight saving time period from April 4th to October 31st, 2004. During daylight saving time period, the GPS time had four-hour difference with local time.

As the final reduction step, second-by-second GPS instrumented vehicle data were aggregated to the STN level. In other words, the STN-based mean travel speed was calculated so that they could be compared with the VDS speeds. The STN-based mean speed was calculated by averaging all GPS data points within a polygon, and the time of estimated speed was determined by the time stamp of the middle point of GPS data points within polygon, which was designated as the reference time of STN. Table 4-1 summarizes the final GPS dataset after data reduction procedure, which will be

matched with VDS dataset for speed comparison. Finally, 138,284 STN-based GPS data points from the 76,470 trips of 209 drivers were generated for the speed comparison with VDS data. While northbound dataset has 66,713 STN-based data points from 38,104 trips of 203 drivers, southbound dataset has 71,571 STN-based data points from 38.366 trips of 198 drivers.

TABLE 4-1 Summary of the GPS Dataset after Data Reduction

	Trips	Drivers	STN-Based Data Points
Northbound	38,104	203	66,713
Southbound	38,366	198	71,571
Total	76,470	209	138,284

4.2 Sample Size of GPS Data

Before using the GPS data, an assessment must be conducted to determine whether the GPS speed data are likely to be representative of population speed on freeways. The sample size of trips and drivers is first examined. Figure 4-2 shows the number of trips and drivers for the sampled GPS data of each STNs in this study area. While STN 4001120 has the maximum GPS trips (1,665 trips) and STN 4000028 has the GPS trip data of maximum drivers (153 drivers), STN 4001102 has only one trip. More specifically, 35 out of 39 STNs on the northbound and 31 out of 36 STNs on the southbound have GPS trip data for more than 800 trips of over 100 drivers. Thus, the number of sample size is large enough to conduct comparisons between GPS and VDS

speed. Among 172 all GPS sample drivers, 105 drivers had more than 100 trips along the corridor in both directions.

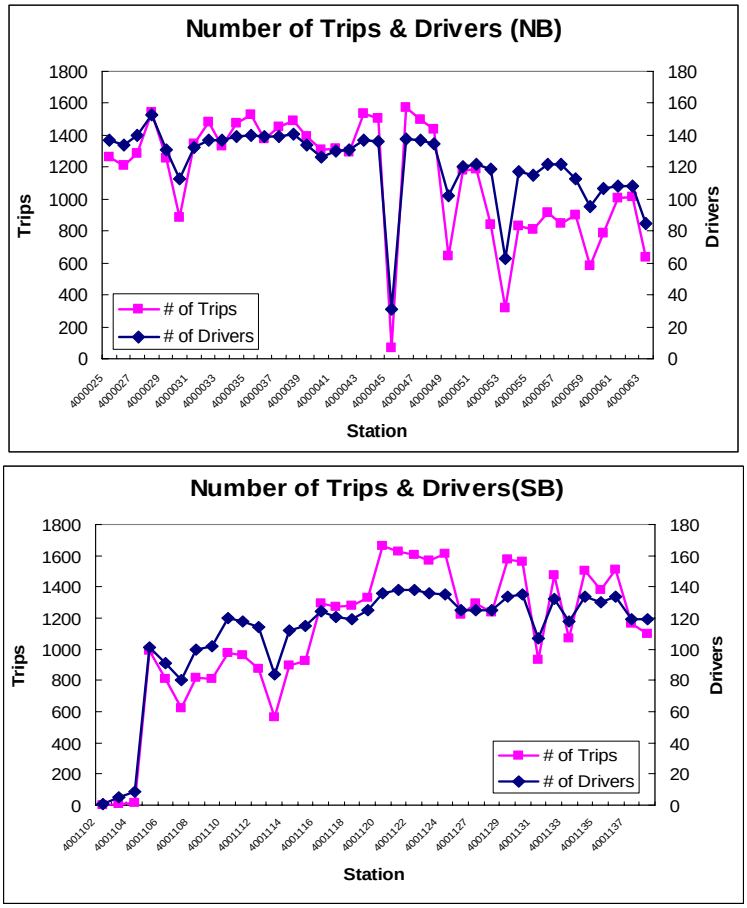


FIGURE 4-2 Number of Trips and Drivers of GPS Data by STNs, Northbound (top) and Southbound (bottom)

Figure 4-3 shows the distribution of GPS speed that consists of 1,017 trips of 130 drivers at LOS A and B. Since the GPS data that obtained nearly at free flow speed follow the normal distribution, GPS data were assumed to be randomly selected for the speed comparisons.

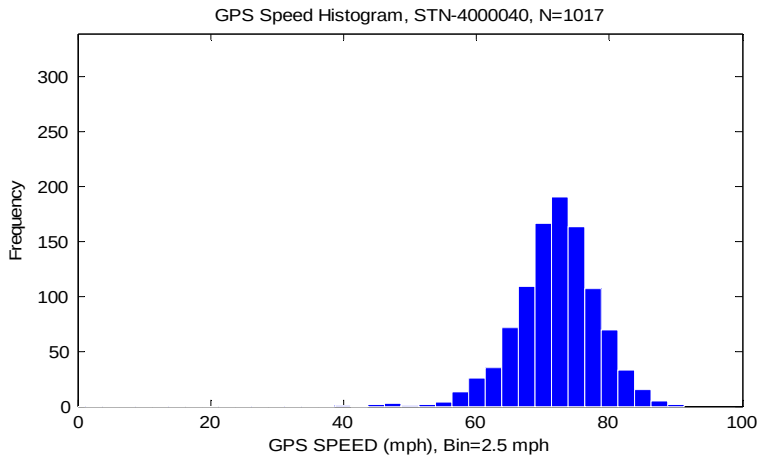


FIGURE 4-3 Distribution of the GPS speed at LOS A to C

4.3 The characteristics of STN-based GPS speed

All GPS speed data aggregated to 20-second intervals to match with the corresponding VDS mean speed, which assumes that the speed of all vehicles within each STN are the same during the 20 seconds. For example, driver "A" (ID 6002902) traveled 0.33 mile segment (STN 4001136) for 19 seconds, thus his average speed is 62.3 mph as shown in Figure 4-4. On the other hand, driver "A" traveled next 0.32 mile segment (STN 4001135) for 23 seconds, and the average speed of his trip on this segment is 51.5 mph. When a driver changes his/her speed dramatically, the speed difference between 20-second intervals can be significant. In addition, since the VDS and GPS

speed data aggregated with 20-second interval by each STN, this study assumes that the geometric characteristics of each STN are homogeneous.

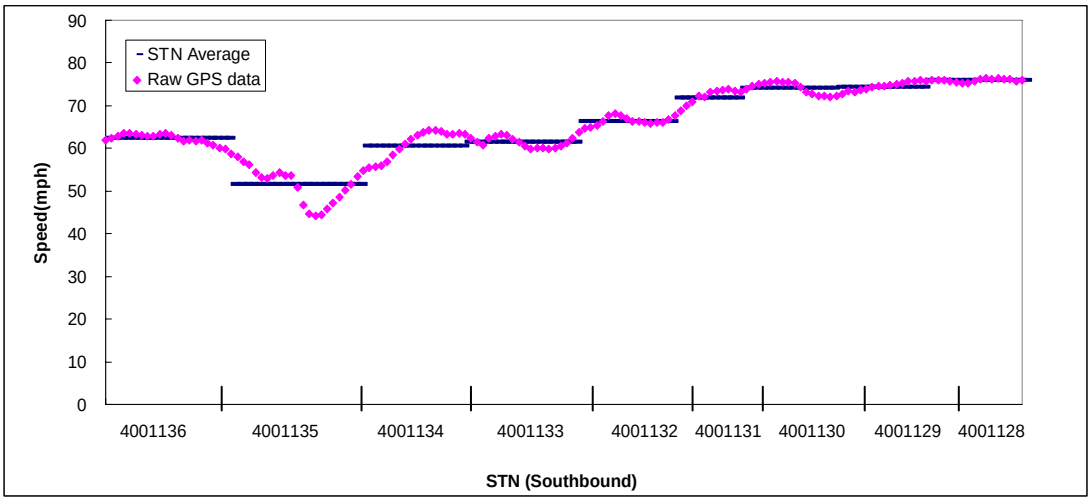


FIGURE 4-4 Original GPS Speed and Aggregated STN-based Speed

In addition to the aggregation characteristics, since the GPS trip data obtained from the individual sample driver, drivers have different individual travel speeds even under same traffic condition, time of day, and location. Thus, GPS speed has greater variance than VDS speed, which will be discussed more detail in next chapter.

4.4 The Lane-by-lane Distribution of GPS Data Points

In addition to the sampling examination of the GPS speed, whether GPS drivers have lane preference among lanes is examined e.g., GPS drivers have tendency to run on the inside or outside of the lane on freeways. The GPS data points of 10 drivers traveled four-lane basic segment during August 2004 of the GA 400 northbound (STN 4000042) were analyzed for lane preference analysis. To determine the lane preference of the

GPS drivers, virtual polygon encompassing whole GPS data points were generated for northbound dataset and then this polygon was divided into two polygons by the centerline of that polygon in GIS application as shown in Figure 4-5. In the polygon process, author assumed that all GPS data points have same random error. The left side of the polygon had 52% of GPS data points as shown in Table 4-2, which means that GPS drivers more likely run on left two lanes than right side two lanes at four-lane segments. However significant difference between left side and right side of the lane did not exist.

TABLE 4-2 Number of GPS Data Points by Sides

	Left Side	Right Side	Total
Data point	2,637	2,516	5,153
Percentage	51.57	48.43	10

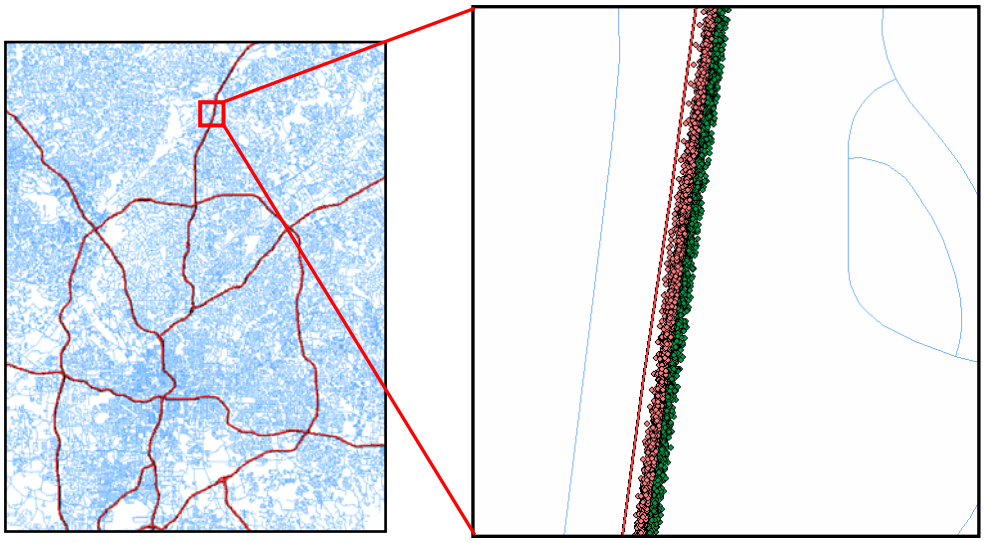


FIGURE 4-5 GPS Trip Data Distribution on the Northbound Freeway

4.5 Curvature and grade effect

Since this study examines the GPS and VDS data of 0.3-mile length STN for the analysis, considering precise curvature and grade within segment was unavailable. Thus, this study assumes that the geometric characteristics of each STN are homogeneous. Ko (2006) examined the effect of curvature and grade in traffic quality measure. He found that grade has significant effect on traffic flow under worse LOS conditions such as D to F, and curvature has significant effect on traffic flow under only LOS A conditions. However, the effects of the grade and curvature should not be significant on this study corridor because the portion of significant grade and curvature area would be the little part of STNs.

4.6 Before and After Removing On/Off-ramp Trips

Even though some stations include on/off-ramps, since the VDS speed only represents the spot speed of all vehicles traveled, VDS speed may not completely capture the on/off-ramp traffic impact. To assess the potential on/off-ramp traffic impact, two GPS speed datasets are created, one before and one after removing on/off-ramp activities. When this study considers only mainline traffic on freeway, all trip data points affected by acceleration and deceleration activities at on/off-ramps were removed from the GPS speed dataset. To remove the impacts of traffic at the on/off-ramp on speed estimates, this study defined about one-mile boundary (three consecutive STNs) from on/off-ramps as an influential segment of the ramps, although HCM 2000 defined 1,500 ft. as an influence area at on/off-ramps. Consequently, the data points of the GPS trip within

one-mile length after entering the on-ramp and before exiting the off-ramp were removed from speed datasets. Figure 4-6 shows the mean speed before and after removing on/off-ramp trips from GPS dataset. Since the GPS and VDS speeds have no difference between before and after removing on/off-ramp data points, this study applies GPS and VDS speed data before removing on/off-ramp data points.

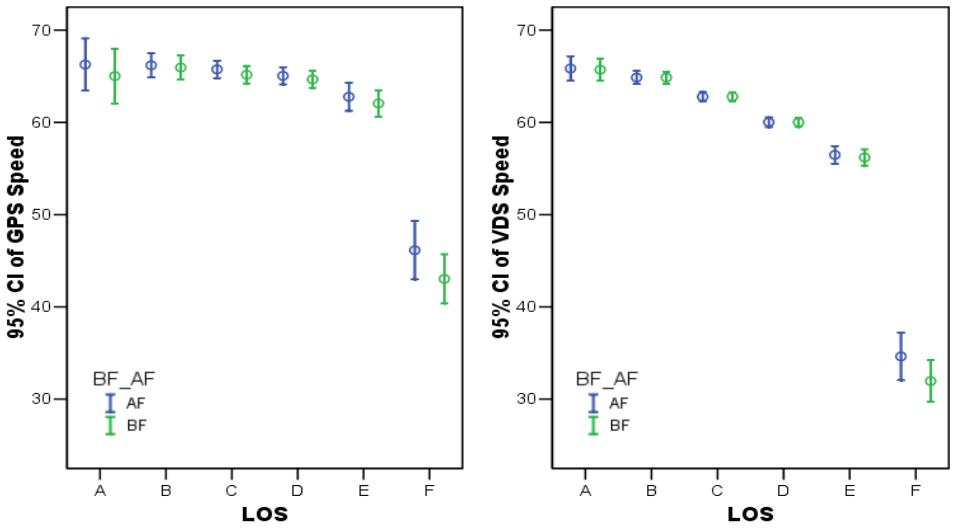


FIGURE 4-6 GPS Mean Speed Before and After Removing On/off Ramp Trips

4.7 Matching GPS and VDS Speed Data

Seven data reduction steps were applied form the data acquisition from individual vehicle to calculation of STN-based mean speed. Initially, 138,284 STN-based GPS data points from 76,470 trips of 209 drivers were generated. Since every STN-based data have own time stamps and STN IDs, the time stamp of VDS speed dataset were searched to match with the time stamp of each GPS trip data. Although initially 138,284 STN-based GPS data points were obtained, the final dataset reduced to

80,370 data points after matching process with VDS dataset. In addition, 178 out of 209 drivers' trip data were selected after matching process between GPS and VDS data.

Table 4-3 summarizes matched dataset for comparisons between GPS and VDS speeds. 2-lane segments have 5,719 data points by 118 drivers, 3-lane segments have 18,182 data points by 169 drivers, and 4-lane segments have 56,469 data points by 173 drivers. Northbound has 38,115 STN-based data points by 156 drivers and southbound has 42,255 STN-based data points by 171 drivers. After matching process, 178 out of 209 drivers' trip data were obtained from initial GPS dataset.

TABLE 4-3 Matched STN-based Dataset for Speed Comparison

		2 lane	3 lane	4 lane	Total
Number of Trips	Northbound	2,357	9,048	26,710	38,115
	Southbound	3,362	9,134	29,759	42,255
	Total	5,719	18,182	56,469	80,370
Number of Drivers	Northbound	107	138	146	156
	Southbound	111	163	164	171
	Total	118	169	173	178

CHAPTER FIVE

COMPARISONS BETWEEN GPS AND VDS SPEED

The goals of comparisons presented in this chapter are to evaluate the accuracy of VDS speed for the purpose of camera calibration and maintenance and to evaluate the differences between speeds collected by STNs and GPS-equipped vehicles as a function of roadways, drivers, and vehicles characteristics. This chapter investigates the speed and the difference between GPS and VDS speed associated with various factors affecting speed differences such as traffic conditions and roadway and environmental characteristics as below.

- Traffic conditions: level of service and truck percentage
- Roadway characteristics: freeway sub-type, number of lanes, and speed limit
- Environmental characteristics: precipitation, daylight, time of day, and weekday/weekend

5.1 Preliminary Analysis for Comparison of GPS and VDS Speeds

Preliminary analysis was first conducted to check the data quality and to understand the basic relationship between the VDS and GPS speed. The initial dataset for this study was collected through on the 79 segments of the GA 400 corridor between January and August 2004. Figure 5-1 shows the average travel speed and the standard deviations of both GPS and VDS data at each STN. A visual examination of data

reveals that STN 4000055 northbound exhibited suspiciously high mean speed compared with the VDS speed of adjacent STNs as well as the GPS speed of the same STN. On the other hand, STN 4001133 southbound has extremely low VDS mean speed compared with other STNs. The VDS speed from the two STNs may be associated with VDS equipment error and/or camera malfunctions. Thus, speed data from the two STNs were assumed to be unreliable and thus excluded from this analysis.

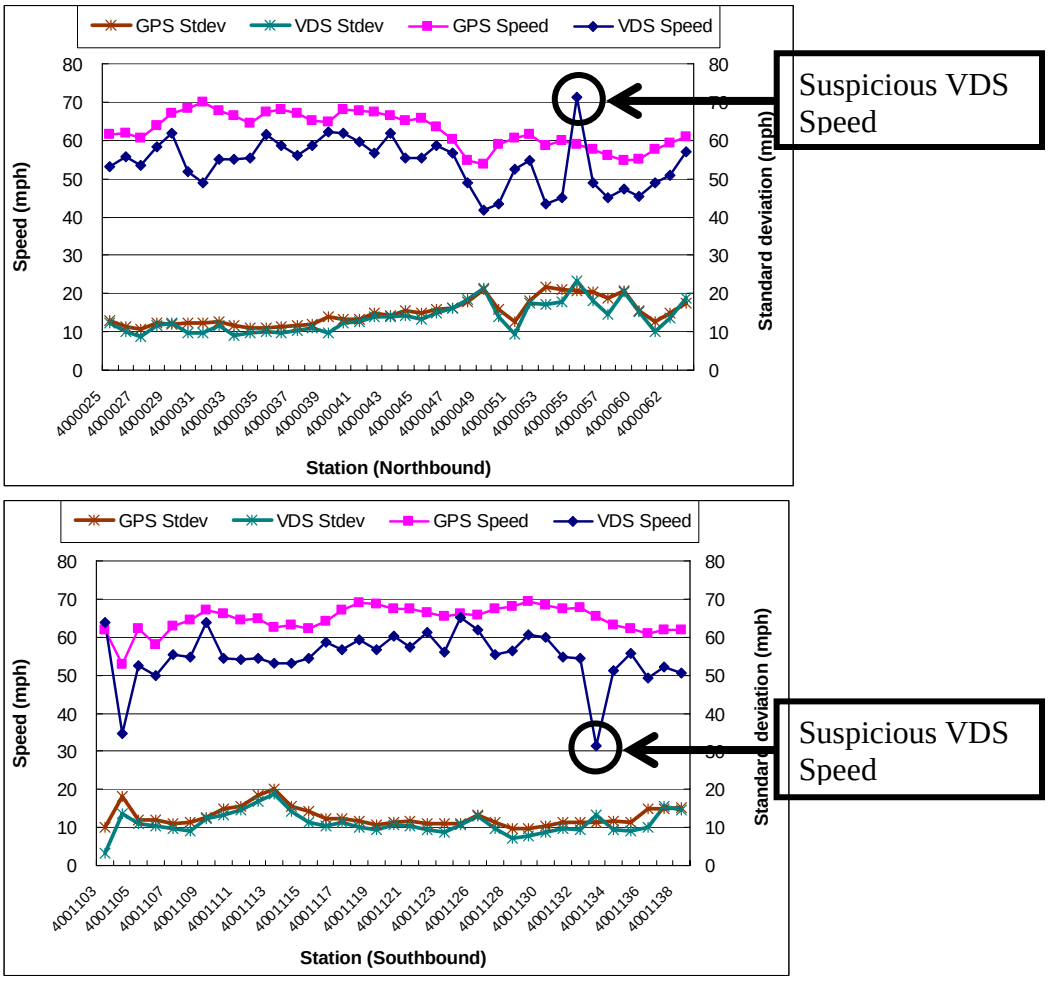
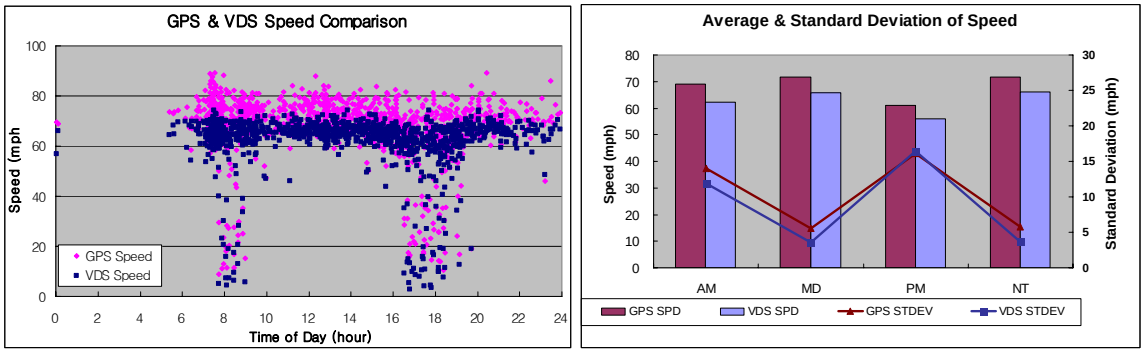


FIGURE 5-1 Visual Examination of STN Data Quality Northbound (top) and Southbound (bottom)

Figure 5-2 shows the raw values of both the GPS and VDS speed by the time of day and the mean speed and the standard deviation of both speeds data under four time periods: AM (6 to 10 am), MD (10 am to 4 pm), PM (4 to 8 pm), and NT (8 pm to 6 am of next day) (right). Raw speed data in Figure 5-2 show that the GPS speed is consistently greater than the VDS speed, except during some parts of AM and PM periods.



Note: Only weekday data were used
AM: 6 AM to 10 AM (4 hrs), MD: 10 AM to 3 PM (5 hrs), PM: 3 PM to 7 PM (4 hrs), NT: 7 PM to 6 AM (11 hrs)

FIGURE 5-2 Raw Speed by Time of Day (left) and Mean and Standard Deviation of GSP and VDS Speed by Four Time Periods (right)

Aggregated data in Figure 5-2 show that GPS speed is 5 to 6 mph higher than VDS speed, and the standard deviations of GPS speed are equal to or greater than those of the VDS during all periods. Both GPS and VDS speed during the AM and PM periods are lower than those during the MD and NT periods, but the standard deviations of both speeds are much higher. During the AM, MD, and NT periods, the standard deviations of the GPS speed are higher than those of the VDS speed. The difference between the standard deviation of GPS and VDS speeds mainly results from the fact that

GPS speed comes from the sample data of GPS-instrumented vehicles, but VDS speed comes from all vehicles detected by VDS cameras. Figure 5-3 shows the scatter plot and histograms corresponding with VDS and GPS speeds. The difference between GPS and VDS speed is significant at high speeds e.g., greater than 50 mph. More specifically, the distribution of the GPS speed has greater variance than that of the VDS speed, and the same characteristic can be seen in Figure 5-2 above. Data points at low speed are most likely obtained during the AM and PM periods.

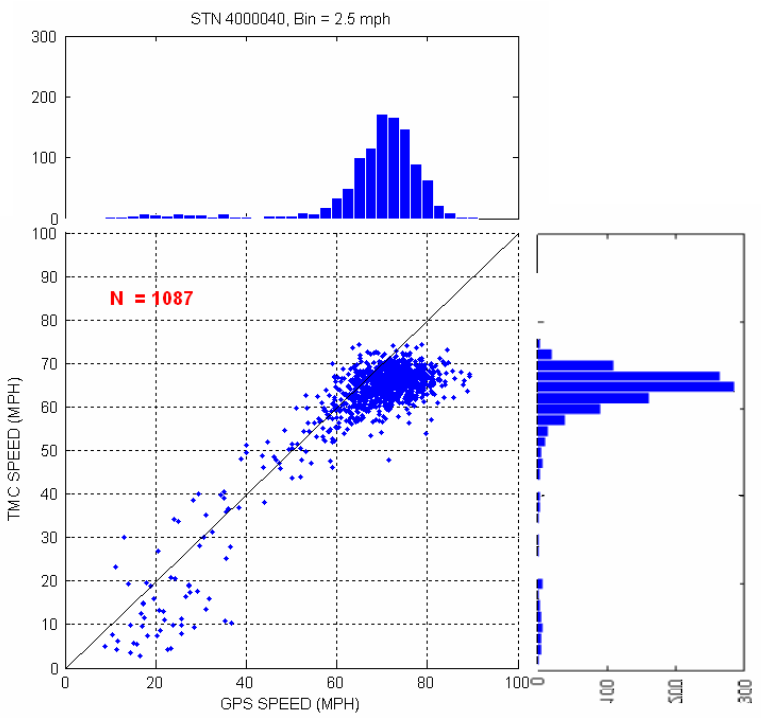


FIGURE 5-3 Scatter Plot and Histograms of GPS and VDS Speed (STN 4000040)

Figure 5-4 shows the histogram of the difference between the GPS and VDS speed, which follows normal distribution. The mean value of the distribution for the speed difference shows that the GPS speed is at an average of 6 mph higher than the VDS speed. The speed difference between the GPS and VDS by the time of day in Figure 5-4 shows the majority of data points located in the negative region for all time periods. GPS verses VDS speed differences during the MD and NT periods are lower than speed difference during AM and PM periods. The cumulative distribution function of the speed difference between the GPS and VDS speed shows that approximately 90% of the VDS speeds are lower than the corresponding GPS speed. Findings from Figure 5-4 demonstrate that the GPS speed is systematically higher than the VDS speed at this station. All other STNs deployed in this research have similar characteristics between GPS and VDS speeds to STN 4000040.

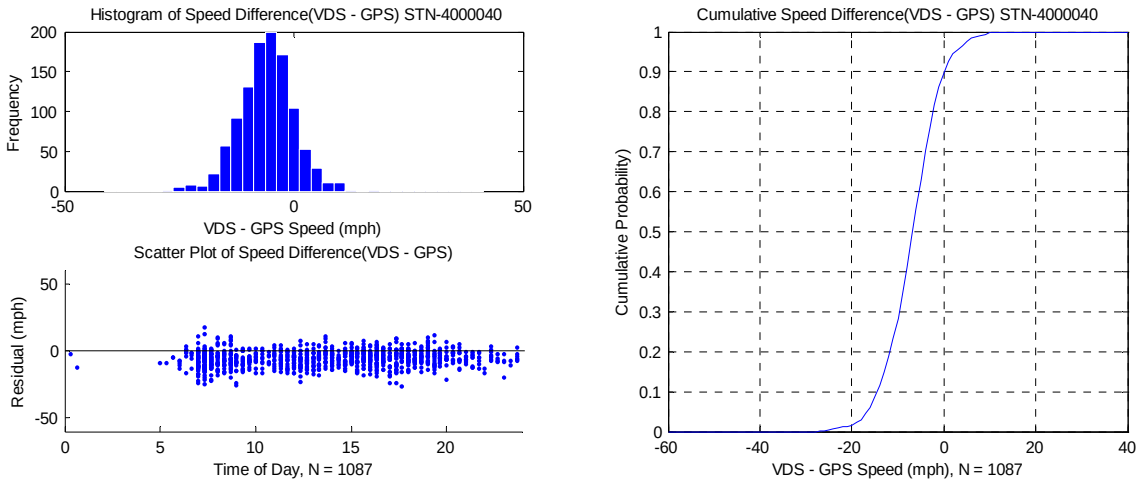


FIGURE 5-4 Speed Difference Between Matched VDS and GPS Data

5.2 Factors Affecting the Difference Between GPS and VDS Speed

This chapter investigates the speed and the difference between the GPS and VDS speed associated with various factors affecting speed difference; traffic conditions and roadway and environmental characteristics. The potential explanatory variable data are categorized and converted into discrete variables. In order to examine the basic characteristics of GPS and VDS speed, this study generates a sub-dataset that accounts for the factors mentioned above and calculates the mean and standard deviations of both GPS and VDS speed. After that, 95% confidence intervals for the mean speed difference are obtained to determine whether the mean speed differences of each group are significant. In addition to the confidence interval analysis, this study applies the Kolmogorov-Smirnov test (KS Test) to examine the magnitude of differences among the distributions at a significance level of 0.05. The Kolmogorov-Smirnov statistics (KS statistics) represent the maximum difference between the two distributions. If KS statistics are greater than the p-value, the test rejects the hypothesis that the two distributions are the not different at the 5% confidence level (Mathworks, 2007). The resulting KS statistics and p-values for the comparisons among the groups are summarized below.

5.2.1 Speed Difference by Traffic Conditions

Level of Service (LOS)

The effects of traffic conditions on the difference between the GPS and VDS speed were investigated first. Information about traffic conditions was obtained from

VDS raw data and converted into level of service A through F by using the traffic density. When converting density into LOS, this study utilized the LOS range defined by *" The Highway Capacity Manual"* (TRB, 2000) as shown in Table 5-1.

TABLE 5-1 LOS Criteria for Freeway Segments

LOS	Density Range (pc/mi/ln) *
A	≤ 11
B	> 11 – 18
C	> 18 – 26
D	> 26 – 35
E	> 35 – 45
F	> 45

* pc/mi/ln indicates passenger cars per mile per lane

The mean and standard deviation of both GPS and VDS speed were calculated for LOS ranges, as shown in Figure 5-5. The GPS speed is greater than VDS speed through all the LOS levels and the difference between the two speeds increases as the traffic congestion (LOS level) worsens. The GPS speed is an average of 11 mph greater than the VDS speed, and the standard deviation of both GPS and VDS speed are very close.

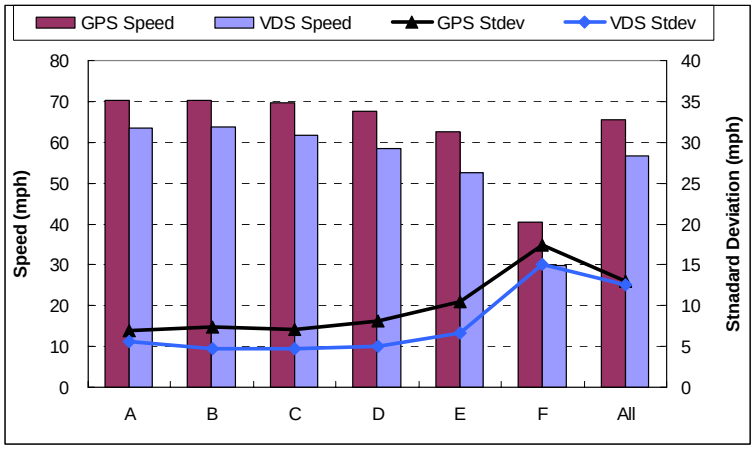


FIGURE 5-5 Mean Speeds and Standard Deviations by LOS

More specifically, the speeds of GPS and VDS data are about 70 mph and 63 mph at respectively the LOS A through C range and decrease to 41 mph and 30 mph at LOS F. However, the standard deviation of the GPS and VDS speed is about 7 mph and 5 mph, respectively, at LOS A through C and increases to 13 mph and 12 mph at LOS F

In addition to the examination of the mean speed, the mean of the difference between GPS and VDS speed was investigated. Figure 5-6 shows the 95% confidence intervals of speed difference between GPS and VDS data for the LOS groups. The figure indicates that the difference between GPS and VDS speed increases as traffic conditions worsen. Thus, as traffic condition worsens, VDS speed accuracy may worsen as previous research (Washington State DOT, 2007). During slower congestion flow traffic, the traffic count and speed accuracies of all non-intrusive detection devises degraded to 10 to 30 mph difference from the baseline system (Middleton and Parker, 2004).

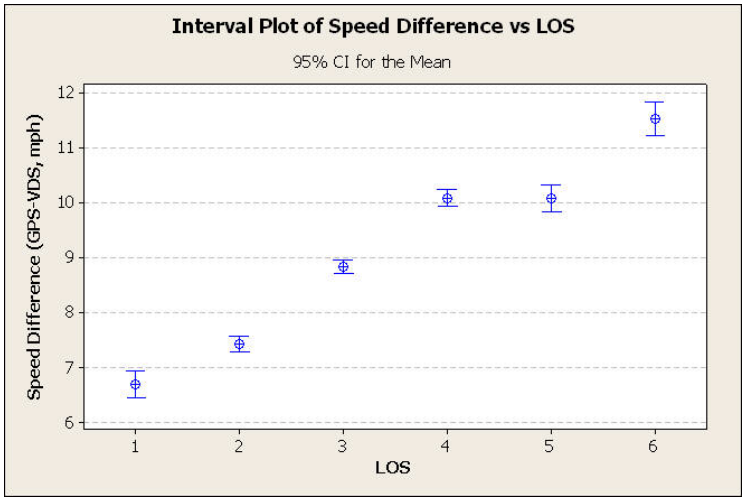


FIGURE 5-6 Confidence Intervals for the Means of Speed Differences by LOS

None of the confidence intervals overlaps except in the LOS D and E ranges, indicating mean speed differences are not statistically different at a significance level of 0.05 across these conditions. Therefore, LOS may significantly influence the difference between GPS and VDS speed

Figure 5-7 shows the distributions of the speed difference between the GPS and VDS data for LOS ranges. Most distributions appear to approach a normal distribution, but the distributions become left-skewed as LOS worsens.

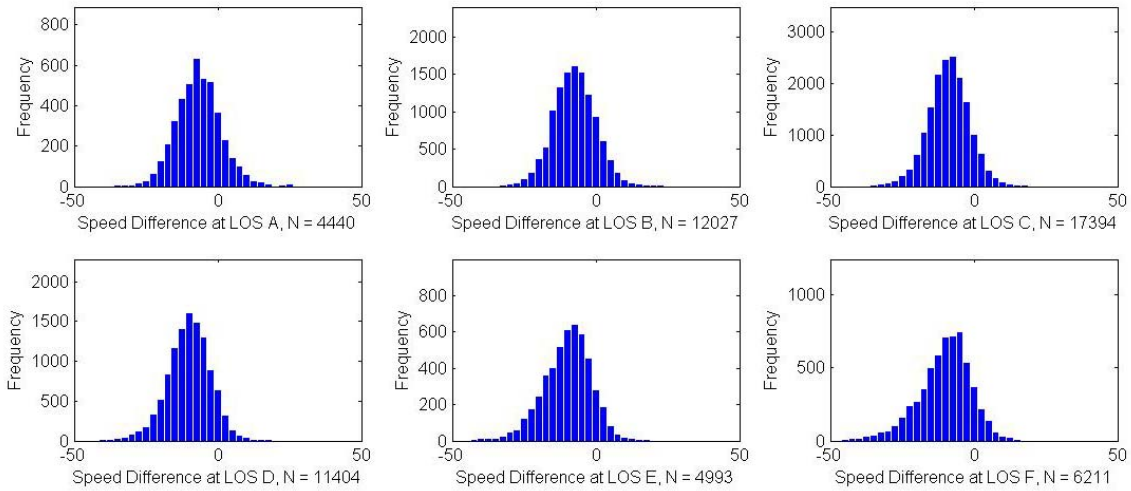


FIGURE 5-7 Distributions of Speed Difference by LOS

Table 5-2 shows the KS test results for the speed difference between the GPS and VDS data among the LOS ranges, indicating all the distributions of all LOS ranges are statistically different at a 0.05 level of significance.

TABLE 5-2 K-S Statistics and P-values for the Pair-wise Comparisons of Speed Difference Distributions by LOS

	A	**B**	**C**	**D**	**E**	**F**
A	-	0.045 (0.000)	0.116 (0.000)	0.169 (0.000)	0.139 (0.000)	0.146 (0.000)
B		-	0.076 (0.000)	0.129 (0.000)	0.111 (0.000)	0.138 (0.000)
C			-	0.066 (0.000)	0.078 (0.000)	0.107 (0.000)
D				-	0.039 (0.000)	0.073 (0.000)
E					-	0.046 (0.000)
F						-

Truck Percentage in Traffic

Hoogendoorn (2005) found that the travel speed of person cars is higher than trucks, and the speed of person cars driving in the left lane is higher than person cars driving in the right lane. The speed of trucks in the left lane tends to higher than trucks in the right lane and the speeds of trucks on either lane are much lower than the person cars (Hoogendoorn, 2005). This study divides heavy truck percentage into five groups as 0 to 3, and 3 to 5, 5 to 7, 7 to 10, and over 10%. To examine the potential truck effect on speed difference under the same conditions, this study uses the data from 4-lane, 65mph speed limit segments. Figure 5-8 shows the mean and the standard deviation of the GPS and VDS data for truck percentage.

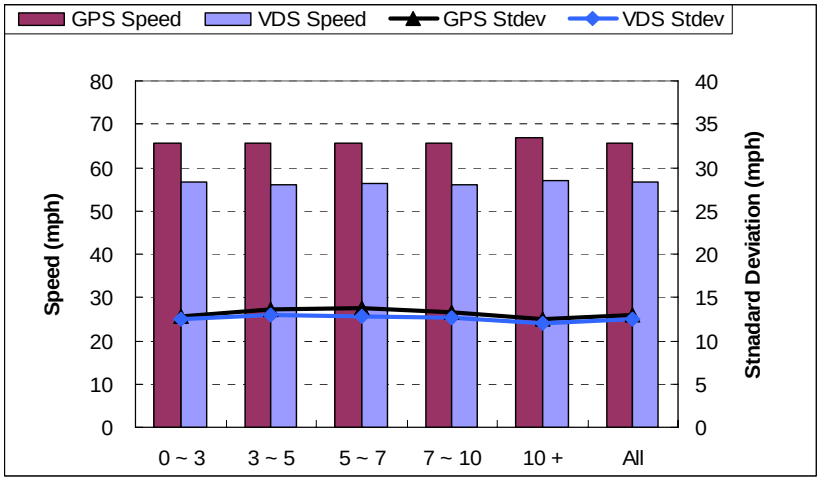


FIGURE 5-8 Mean speed and Standard Deviations by Truck Percentage

The speeds and standard deviations of both GPS and VDS data have same patterns though the truck percentage range. Even though truck percentage increases, the speed and the standard deviation of both GPS and VDS data remain the same. That means the truck percentage does not affect the different between GPS and VDS speed for

this freeway sub-type. However, there may be interaction affects. Grant et al. (1999) found the significant effects on traffic and VDS accuracy.

Figure 5-9 shows 95% confidence intervals of speed difference between the GPS and VDS data for the truck percentage groups. The figure indicates that the speed difference is lowest at the truck percentage of less than 3. As truck percentage increases, the difference between GPS and VDS speed increases and the confidence intervals also increase. Although confidence interval has a pattern as truck percentage increases, most of them overlap each other. However, the speed difference between two groups less than 3% and 3% to 5% of truck percentage are significantly different. Further, when a dataset is divided by two groups as less than 3% and greater than 3% or equal to 3% range, the confidence intervals of two groups do not overlap. Thus, the mean speed differences of two groups (less than 3% and 3% to 5%) are statistically different at a significance level of 0.05, and may significantly contribute to the difference between the GPS and VDS speed.

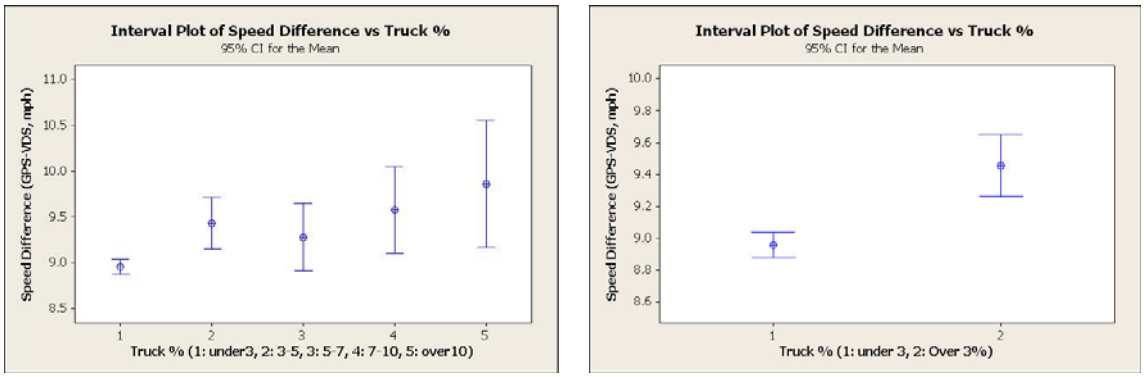


FIGURE 5-9 Confidence Intervals for the Mean of Speed Difference by Truck Percentage of Five Groups (left) and Two Groups (right)

Figure 5-10 shows the speed distributions of the GPS and VDS data for the truck percentage groups. The distributions of the three groups (less than 3%, 3% to 5%, and 5% to 7%) appear to be close to the normal distribution, but the last two groups, 7 to 10 and greater than 10% groups do not. The KS test results for the speed difference by truck percentage, indicate that the distributions of all groups are statistically different at a significance level of 0.05 as shown in Appendix A.

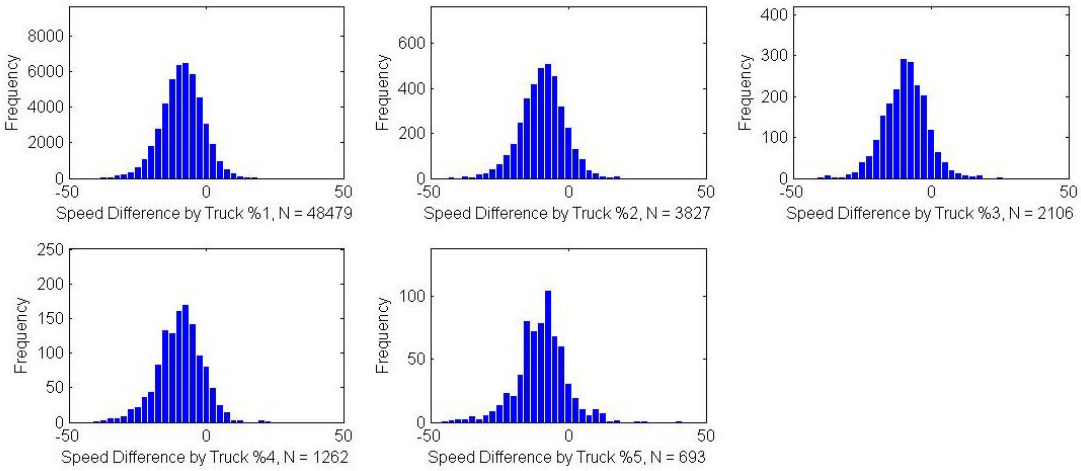


FIGURE 5-10 Distributions of Speed Difference by Truck Percentage

5.2.2 Speed Difference by Roadway Characteristics

Roadway characteristics, including number of lanes, speed limit, and freeway sub-type are also investigated in order to examine factors affecting the difference between GPS and VDS speed.

Number of Lanes

The study corridor consists of 10 two-lane, 22 three-lane, and 45 four-lane segments. The four-lane segments consists of segments subject to either 55 mph or 65 mph the speed limits. This study considers only 65 mph speed limit segments to investigate the potential effect of the number of lanes. Thus, the 18 STNs of the 55mph speed limit are excluded in this examination. Figure 5-11 shows the mean and the standard deviation of the GPS and VDS data for number-of-lane groups. The GPS speed is greater than the VDS speed for all lane groups, and the mean speeds of both GPS and VDS data increase as number of lane increases. However, the standard deviations of both speeds are very similar to each other except lane 2 group.

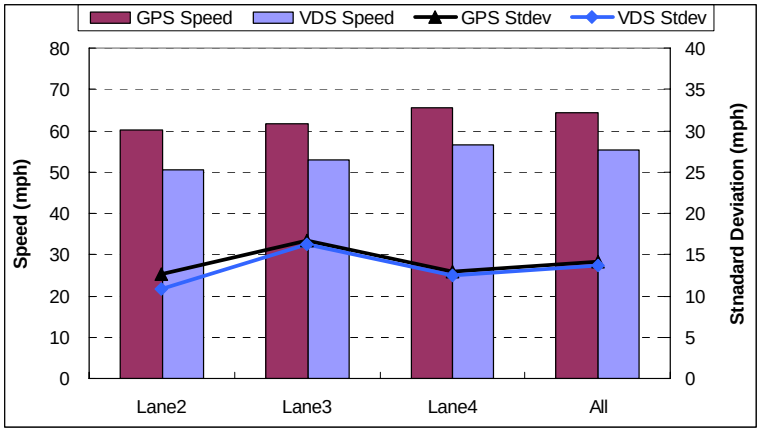


FIGURE 5-11 Mean Speeds and Standard Deviations by Number of Lanes

Figure 5-12 shows 95% confidence intervals of difference between the GPS and VDS speed data for the lane groups. The figure indicates that the difference between the GPS and VDS speed increase as the number of lanes increases, and the magnitude of

confidence interval decreases as the number of lane increases. Because none of the confidence intervals overlap, the difference of the mean speed among the three lane groups is statistically different at the 0.05 significance level. Therefore, number of lanes may significantly contribute to the difference between GPS and VDS speed, which means that as number of lanes increases, VDS data accuracy may decrease as previous research(Grant et al., 1999) or the sampling bias of the GPS speeds may increase. The 2-lane segment has wider confidence interval than 3-lane, and 4-lane segment due to the much small number of dataset. The number of dataset for each segment is 1,943 for 2-lane, 18,182 for 3-lane, and 56,469 for 4-lane segment as shown in Figure 5-13.

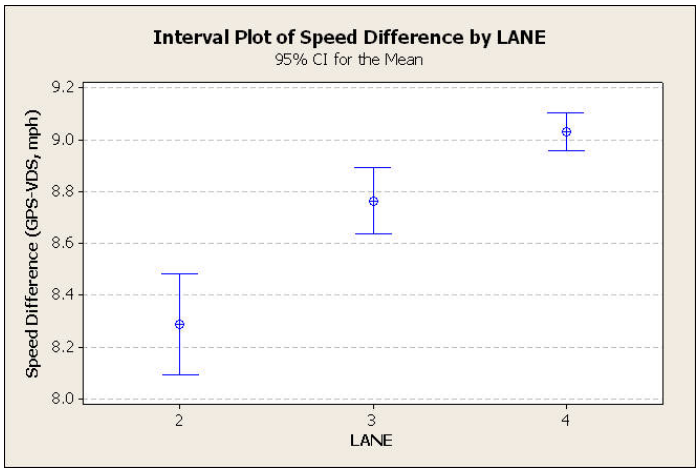


FIGURE 5-12 Confidence Intervals for the Means of Speed Differences by Number of Lanes

Figure 5-13 shows the speed distributions of GPS and VDS data for number of lanes. The distributions of the three-lane and four-lane group appear to be close to the normal distributions. The KS test results for the speed difference between the GPS and VDS data among number-of-lane groups indicate that all the distributions of all groups are statistically different at a level of significance 0.05 as shown in Appendix A.

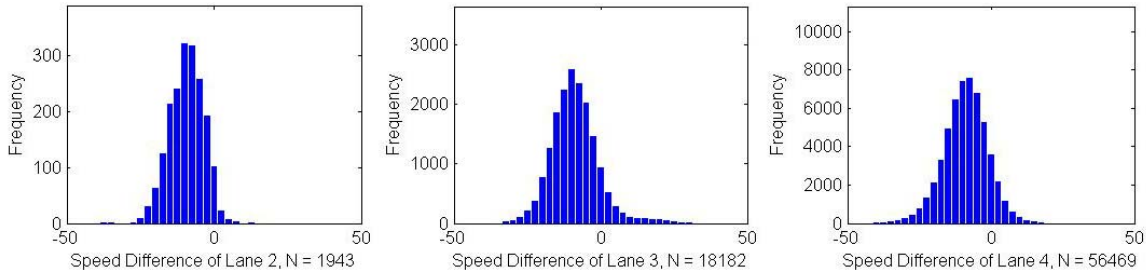


FIGURE 5-13 Distributions of Speed Differences by Number of Lanes

Speed Limit

The study corridor consists of 20 55-mph and 58 65-mph segments. The southern part of the study corridor has a 55-mph speed limit, and the speed limit continues along the ten consecutive STNs for about the three-mile-length corridor. Then the speed limit increases to 65 mph on the remaining segments of the corridor. To examine the effect of the speed limit, this study considers only the four lane segments; 16 STNs of 55 mph and 27 STNs of 65 mph. Figure 5-14 shows the mean and the standard deviation of the GPS and VDS data for the speed limit.

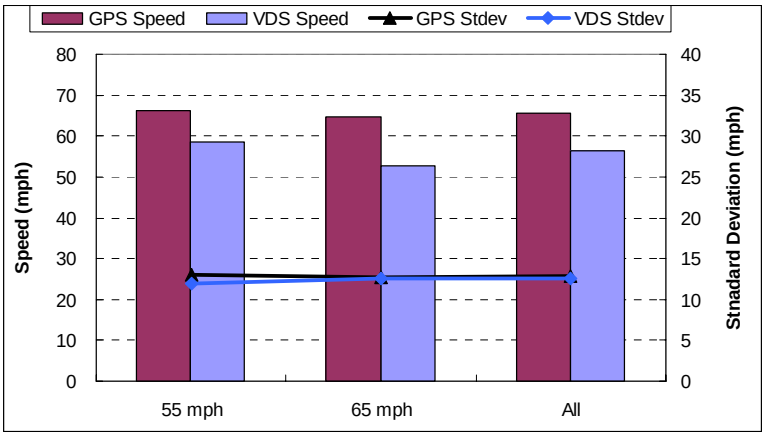


FIGURE 5-14 Mean Speeds and Standard Deviations by Speed Limit

The GPS speed on the 55-mph speed limit segments was 66.1 mph and decreased by 1.5 mph on the 65 mph segments. However, the VDS speed on the 55 mph speed limit segments was 58.6 mph and decreased by 5.8 mph on the 65 mph segments. That is although the speed limit increases from 55 mph to 65 mph, the mean speed of both GPS and VDS data decrease. The facility characteristics of the corridor possibly explain this unexpected finding. Even though the 55 mph speed limit segments interchange with I-285, vehicles on these segments do not experience heavy traffic conflicts. However, the 65 mph speed limit segments have five on/off-ramps and number of lane decreases to three. Thus, the significant conflicts among vehicles that occurred within 65 mph speed limit segments may yield lower mean speed than within the 55 mph segments.

Figure 5-15 shows 95% confidence intervals of speed difference between the GPS and VDS data for the two speed limit groups. The speed difference from the 55 mph speed limit group is greater than that of the 65 mph speed limit group. Since two

confidence intervals do not overlap, the mean speed difference between 55mph and 65 mph speed limit group are statistically different. Thus, the speed limit variable may contribute to the difference between GPS and VDS speed. However, the speed limit effects are probably confounded by design and operational parameters and should not be relied upon. That is there is a large significant difference, but it is probably not actually related to speed limit. The speed difference between two groups was more affected by geometric characteristics and traffic condition than number of lanes themselves. Further study for the combination between number of lanes and other variables is required.

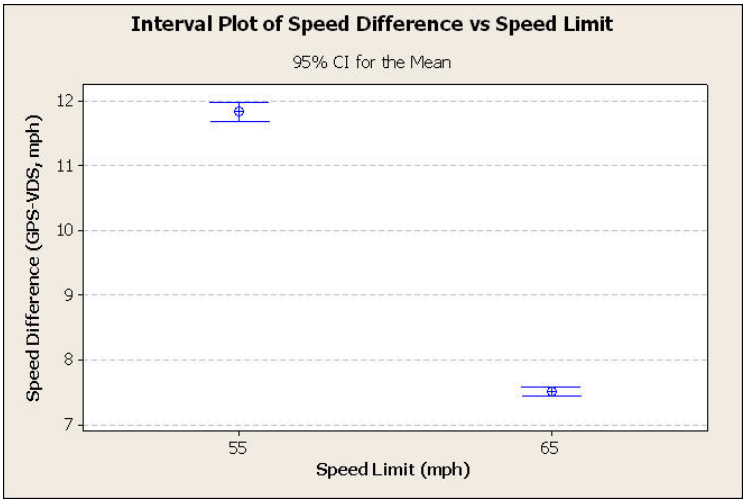


FIGURE 5-15 Confidence Intervals for the Means of Speed Differences by Speed Limit

Figure 5-16 shows the speed distributions of the GPS and VDS data for speed limit groups. Both distributions appear to approach the normal distribution. The distribution of the 55 mph speed limit group has greater variance than that of the 65 mph speed limit group. The KS test results for the speed difference between GPS and VDS data indicate that statistically two different distributions are at a significance level of 0.05 as shown in Appendix A.

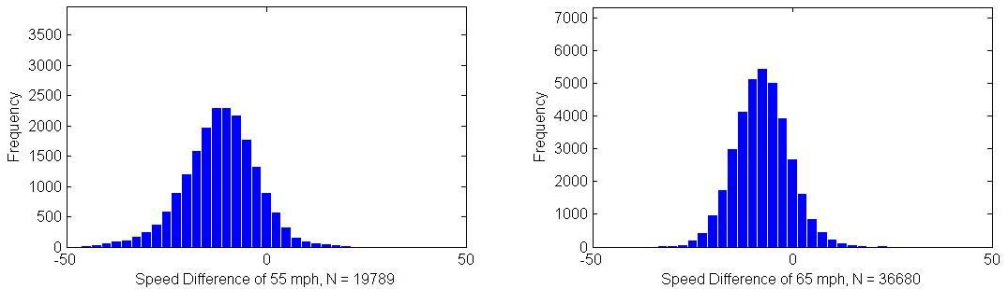


FIGURE 5-16 Distributions of Speed Difference by Speed Limit

Freeway Sub-type

This study investigates the effect of freeway sub-type on the speed difference between GPS and VDS data. Although the study corridor has many on/off-ramps of 10 two-lane, 22 three-lane, and 45 four-lane segments, only four-lanes segments are considered for investigating the effect of freeway sub-type to compare with other factors. Figure 5-17 shows the mean and standard deviation of the GPS and VDS data for the freeway facility sub-types. GPS speed is greater than VDS speed for the all freeway sub-type groups. The GPS speeds are very similar both basic and on-ramp segments, but the VDS speed of basic segments is less than on-ramp segments. The mean speed

of the off-ramp sub-type is lowest among three freeway sub-types, which appear reasonable for the spill back effect of the off-ramp.

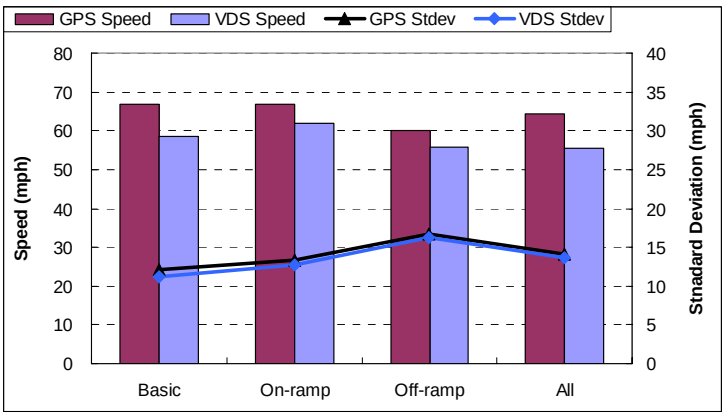


FIGURE 5-17 Mean Speeds and Standard Deviations by Freeway Sub-type

Figure 5-18 shows 95% confidence intervals of the speed difference between GPS and VDS data for freeway sub-type groups. The figure indicates that the difference between GPS and VDS speed is the largest for basic segments and smallest for off-ramp segments. Therefore, the findings indicate that the freeway sub-type variable may contribute to the difference between GPS and VDS speed. This finding appears reasonable because the vehicles of off-ramp segment tend to decrease their speed to change lanes or to stop their vehicles at the first intersection on the surface road connected to the ramp. Because none of the confidence intervals overlap, the mean speed differences across the three freeway sub-type groups are statistically different at a significant level of 0.05. Therefore, the effect of freeway sub-type may significantly contribute to the difference between GPS and VDS speed

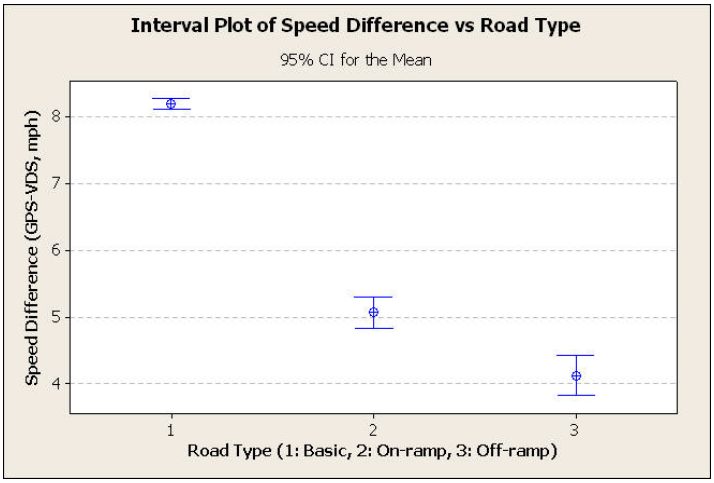


FIGURE 5-18 Confidence Intervals for the Means of Speed Differences by Freeway Sub-type

Figure 5-19 shows the speed distributions of the GPS and VDS data for the freeway sub-type groups. The distributions of basic and on-ramp segments appear to approach the normal distribution. The KS test results for the speed difference between the GPS and VDS data among the freeway sub-type groups indicate that the distributions of all groups are statistically different at a 0.05 level of significance.

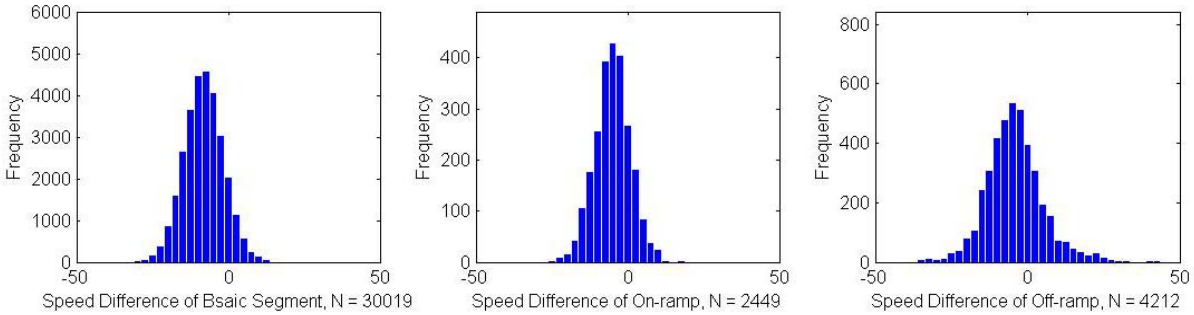


FIGURE 5-19 Distributions of the Speed Difference by Freeway Sub-type

5.2.3 Speed Difference by Environmental Characteristics

As mentioned earlier, this study utilized the precipitation data from the National Oceanic & Atmospheric Administration (NOAA) to consider the potential effects of pavement conditions, visibility, and sunrise and sunset time on the difference between the GPS and VDS speed.

Weather Condition

The NOAA data indicate that 69 out of 244 days from January to August have positive precipitation data. Figure 5-20 shows the mean and the standard deviation of the GPS and VDS data for the weather groups. The GPS speed is greater than the VDS speed among the two weather groups, and mean speed of clear-day group is higher than inclement-day group both GPS and VDS.

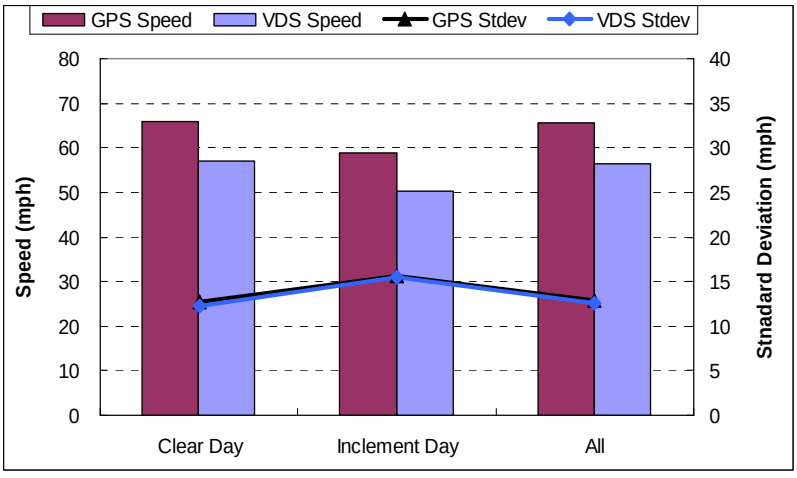


FIGURE 5-20 Mean Speeds and Standard Deviations by Weather

Figure 5-21 shows 95% confidence intervals of the speed difference between the GPS and VDS data for the weather groups. The figure indicates that the difference between the GPS and VDS speed during non-inclement periods is little higher than during inclement periods, but the magnitude of the interval of the inclement period is much greater than that of the clear-day periods. The confidence interval of inclement days is much wider than that of clear days due to the small number of dataset (N = 3,912) compared to the clear-day periods (N = 76,458) as shown in Figure 5-22. Another reason of wider confidence intervals of inclement days may be due to higher speed variability during inclement weather.

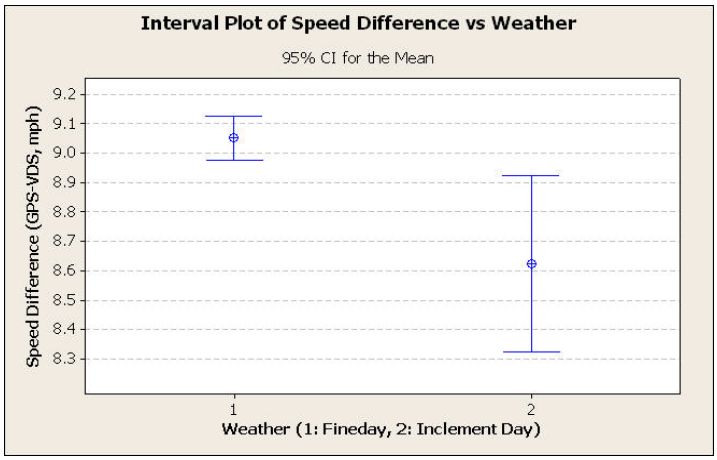


FIGURE 5-21 Confidence Intervals for the Means of Speed Differences by Weather

Since the two confidence intervals do not overlap, the mean speed differences between the two groups are statistically different at a significance level of 0.05. Therefore, variable weather may significantly contribute to the difference between the GPS and VDS speed. In addition, Figure 5-22 shows the speed distributions of the GPS and VDS data for the weather groups. Both distributions appear to be close to the normal

distribution. The KS test results for the speed difference between the GPS and VDS data between the weather groups indicate that the two distributions are statistically different at a 0.05 level of significance as shown in Appendix A.

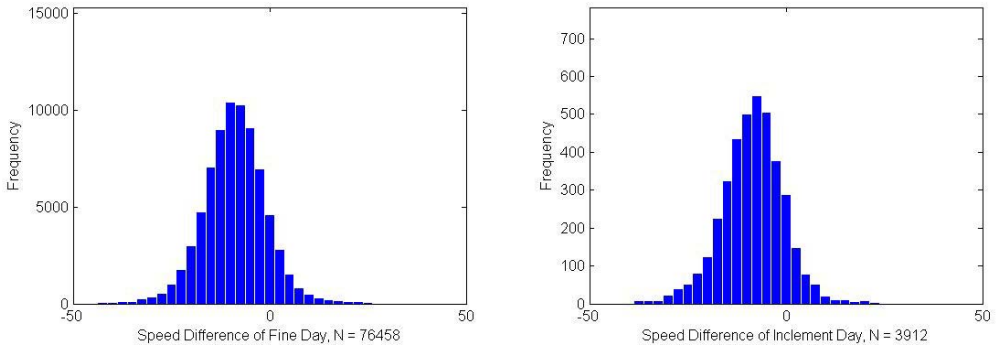


FIGURE 5-22 Distributions of Speed Difference by Weather

Daylight/Darkness

Because VDS cameras suffer from lens glare, which is potentially significant at sunrise and sunset, 30 minutes before sunset and 30 minutes after sunrise are considered twilight time periods. Figure 5-23 shows the mean and the standard deviation of the GPS and VDS data for the day time, nighttime, and twilight time periods.

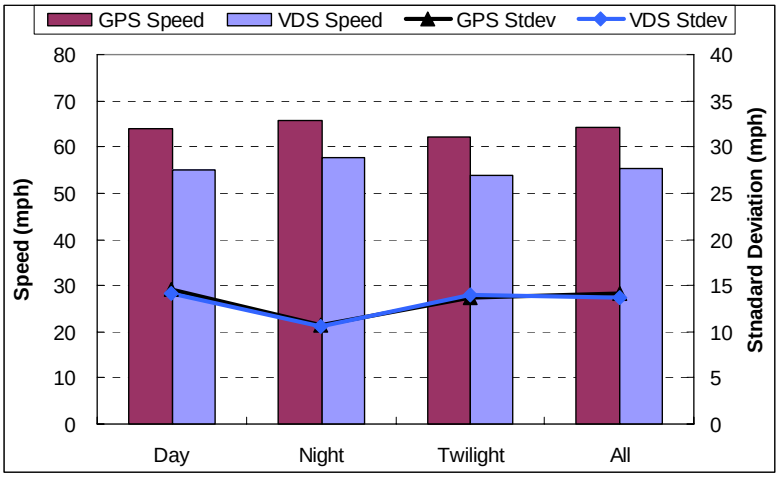


FIGURE 5-23 Mean Speeds and Standard Deviations by Daylight

Although the GPS speeds during three time periods are very similar, GPS speed is highest as 65.8 mph at night time, but lowest as 62.1 mph at twilight time period. The reason why the speeds at night are higher than those of during the day and twilight time periods is that more congested traffic generally occurred during daytime rather than at nighttime. Thus, all vehicles at night hardly experienced congestion compared to the vehicles during the day and twilight time periods. In addition, the mean speed during twilight time period is lower than day time period, which means that twilight affects driver's speed choice.

Figure 5-24 shows 95% confidence intervals of the speed difference among the GPS and VDS data for day, night, and twilight time groups. The speed difference of the non-inclement daytime group is higher than the nighttime group. However, speed difference at twilight time period has very wider confidence interval than those of day and night time periods due to the much small number of dataset as shown in Figure 5-25.

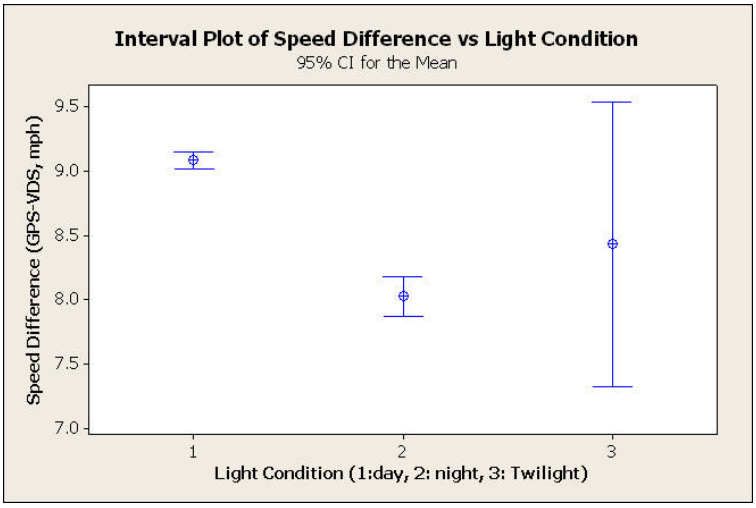


FIGURE 5-24 Confidence Intervals for the Means of Speed Differences by Daylight

Because the two confidence intervals of day and night time periods are not overlap, the mean speed difference between daytime and nighttime are statistically different. Therefore, variable daylight may significantly contribute to the difference between GPS and VDS speed. In contrast, since the confidence interval of twilight time period overlaps with the confidence intervals of the day and night time periods, this study did not yield the significant effect of twilight on speed difference as found by previous research (Hori, 1997; Klein, 1993). Figure 5-25 shows the speed distributions of GPS and VDS data for the day and night groups. Both the distributions during day and night time periods appear to approach to a normal distribution, but the distribution of twilight time period does not due to the small number of dataset. The KS test results for the speed difference between the GPS and VDS data in Appendix A indicate that three distributions are statistically different at a significance level of 0.05.

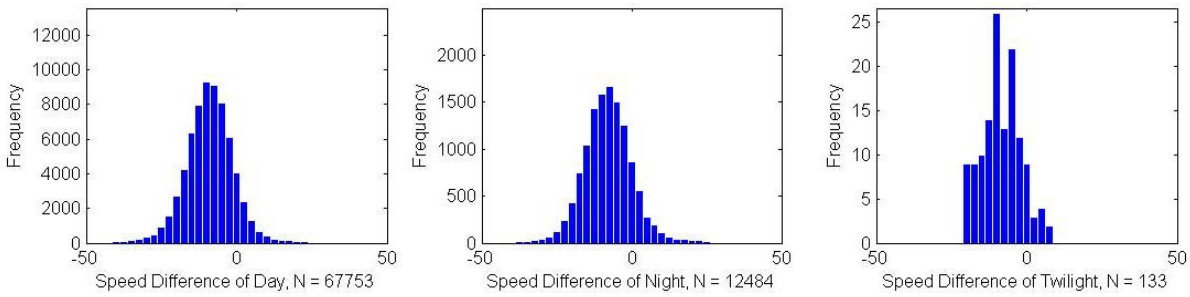


FIGURE 5-25 Distributions of the Speed Difference by Daylight

Time of Day

This study defines the four different day time periods, AM (6 am to 10 am), MD (10 am to 3pm), PM (3 pm to 7 pm), and NT (7 pm to 6 am of next day). Numerous commute trips that occur during peak periods have directional characteristics i.e., most home-to-work trips are southbound during the am peak period, and most work-to-home trips are southbound during the pm peak period of the GA 400 corridor. In order to avoid the conflict of traffic characteristics between AM and PM periods, this study used only southbound dataset in this step. Figure 5-26 shows the mean and the standard deviation of both GPS and VDS speed during the four different time periods.

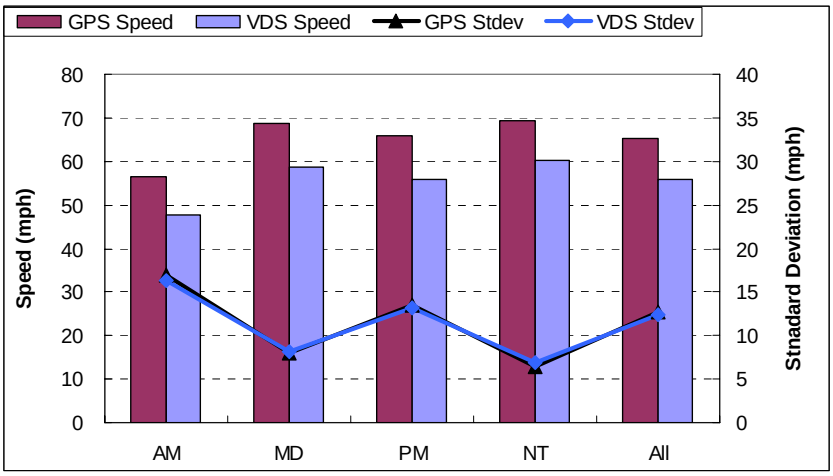


FIGURE 5-26 Mean Speeds and Standard Deviations by Time of Day

GPS speed is greater than VDS speed during all time periods as same as the LOS case. During the four time periods, the AM period has the lowest mean speed of both GPS and VDS data on southbound because commute trips from home to work place dominate during the AM period than during any other periods, thus drivers experienced the most congested traffic condition during AM period. During the congested periods such as AM and PM periods, travel speed tends to decrease and the standard deviation of the speed tends to increase. During the MD and NT periods, the mean of GPS and VDS speed are almost at 69 mph close to the free flow speed, and both speeds have the lowest standard deviations.

Figure 5-27 shows 95% confidence intervals of the speed difference between the GPS and VDS data for time of day groups. AM period has the lowest mean speed among the four time periods in Figure 5-26 but has the highest speed difference between GPS and VDS data among four time periods in Figure 5-27.

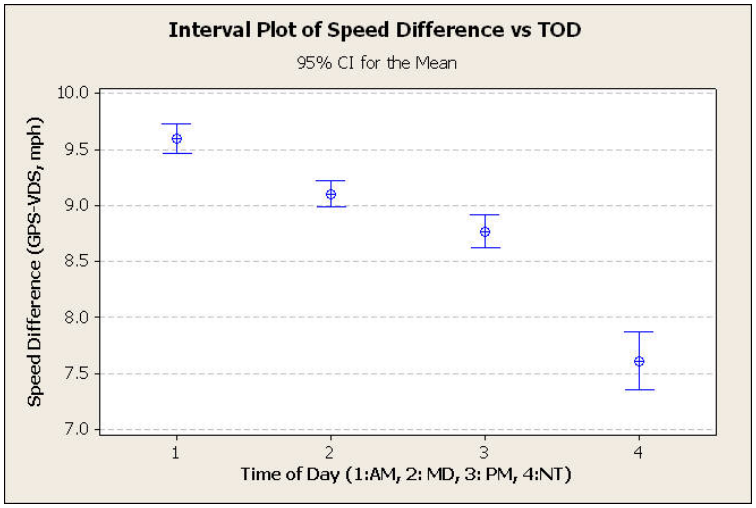


FIGURE 5-27 Confidence Intervals for the Means of Speed Differences by Time of Day

The situation (the mean speeds of both GPS and VDS data are low but speed difference between GPS and VDS data is high) shows the same results as those in the above the LOS case. In other words, as congestion level increases, VDS speed accuracy may decrease. Because none of the confidence intervals overlap, the mean speed differences among four time periods are statistically different at a significance level of 0.05. Therefore, the variable time of day may significantly contribute to the between GPS and VDS speed.

Figure 5-28 shows the speed distributions of the GPS and VDS data for the four time-of-day groups. All distributions appear to be close to the normal distribution. The KS test results for the speed difference between the GPS and VDS data for the time-

of-day ranges in Appendix A indicate that all the distributions of all the time-of-day ranges are statistically different at a 5% level of significance.

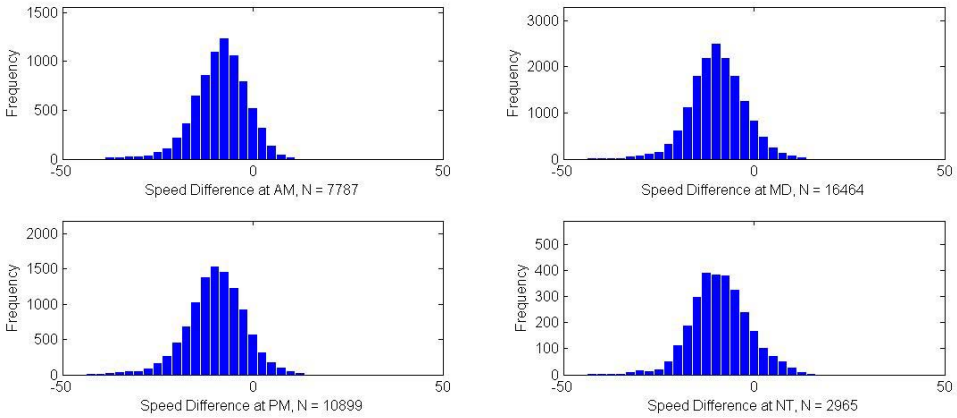


FIGURE 5 -28 Distributions of Speed Difference by Time of Day

Weekdays/Weekends

The difference between weekdays and weekends is also examined. All trips are occurred during Saturdays, Sundays, and official holidays are defined as weekend trips, and 74 out of 244 days are categorized as weekend days. The mean and the standard deviation of GPS and VDS data were calculated as Figure 5-29.

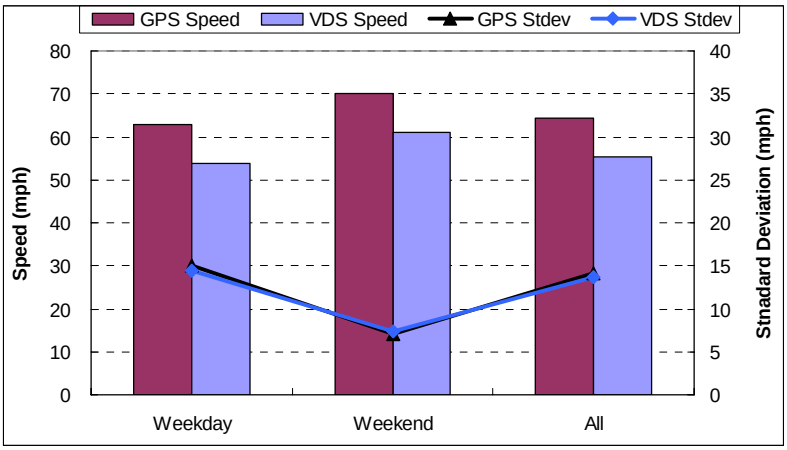


FIGURE 5-29 Mean Speeds and Standard Deviations by Weekday

The GPS speed is greater than VDS speed for both weekday and weekend periods. However, the standard deviations of both speeds are very similar. The mean speed on weekdays is lower than weekends, the standard deviation of the speed on weekdays is higher than during weekends. These findings appear to be reasonable because the level of congestion on freeways during weekends is much lower than that of weekdays. Figure 5-30 shows 95% confidence intervals of the speed difference between GPS and VDS data for the weekday/weekend groups. The speed differences are very similar as about 9.0 mph of both weekdays and weekends. Since the two confidence intervals overlap, the mean speed difference between weekday and weekend is not statistically different. Therefore, the variable weekday/weekend may not significantly contribute to the difference between the GPS and VDS speed.

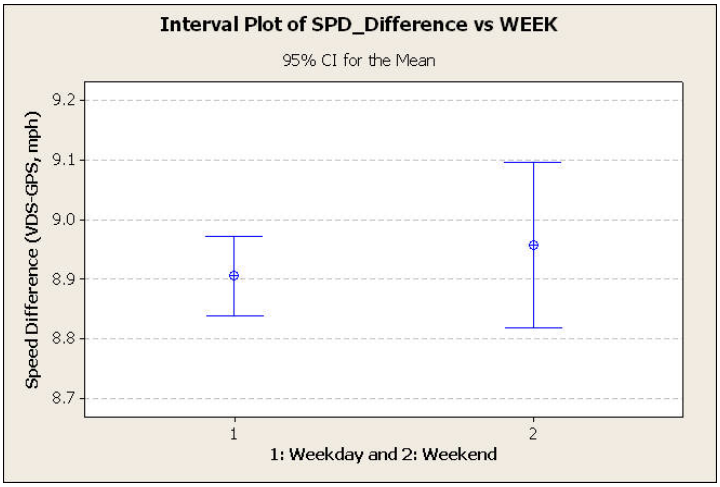


FIGURE 5-30 Confidence Intervals for the Means of Speed Differences by Weekday

Figure 5-51 shows the distribution of the speed difference between GPS and VDS data for the weekday and weekend groups. Both the distributions appear to close to the normal distribution. The KS test results for the speed difference between GPS and VDS data in Appendix indicate that two distributions are statistically different at a significance level of 0.05. However, the p-value 0.03 in parentheses of suggests that the two distributions may not be statistically different at a higher significance level (i.e., 0.01).

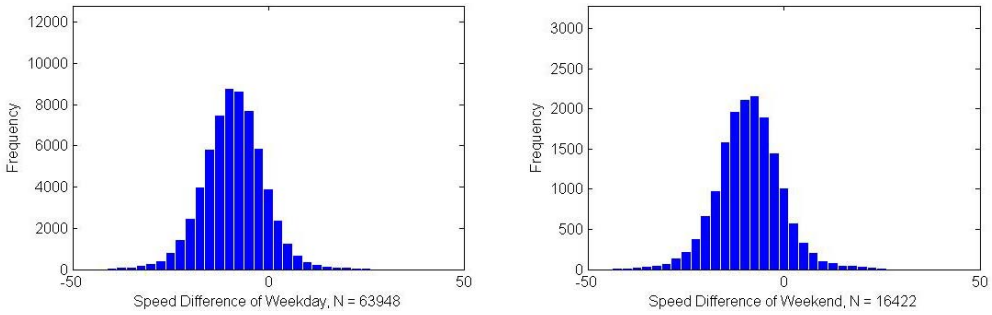


FIGURE 5-31 Distributions of Speed Difference by Weekday

5.3 Summary

This chapter examined nine factors independently to identify factors significantly contribute to the difference between GPS and VDS speed. Groups for each factor were identified and various analyses, including the mean speed comparisons, confidence intervals, the distributions of speed differences, and KS tests were conducted. The following is a list of the findings of this study.

- The speed difference analysis with traffic conditions shows that LOS and truck percentage in traffic may contribute to the difference between GPS and VDS speed.
- The speed difference analysis with roadway conditions shows that the number of lanes, free-flow speed, and freeway sub-types may contribute to the difference between GPS and VDS speed.
- The speed difference analysis with environmental characteristics shows that inclement weather, daylight, and time of day may contribute to the difference between GPS and VDS speed.

The results of this analysis found that eight factors significantly contribute to the difference between GPS and VDS speed, and VDS speed accuracy may be affected by these same eight factors. However, these eight factors need to be considered simultaneously to consider interactions and confounding effects among factors. For example, roadway characteristics, such as number of lanes, the speed limit, and the freeway sub-type are closely related together for the speed. Thus, in the following

chapter, this study will examine how relationships among these eight factors contribute to differences between GPS and VDS speed.

CHAPTER SIX

CLASSCIFICATION AND REGRESSION TREE ANALYSIS FOR THE DIFFERENCE BETWEEN GPS AND VDS SPEED

6.1 Background

By using explanatory variables chosen as significantly affect to the difference between GPS and VDS speed in the previous chapter, this chapter utilizes the classification and regression tree (Recarte and Nunes) analysis technique to identify explanatory variables maximizing measure of difference in the dataset. This technique, referred to as binary recursive partitioning or hierarchical tree-based regression (HTBR) techniques, is similar to forward stepwise variable selection methods (Hallmark, 1999). This tree-building process continues iteratively to answer the following questions: 1) *which variable among all independent variables in the model can produce the maximum reduction in variability of the dependent variable and* 2) *which value of the selected variable (discrete or continuous) can generate the maximum reduction in variability of the response?*.(Washington, 2000; Wolf et al., 1998) In other words, trees explain the variation of a single response variable by repeatedly splitting the data into more homogeneous groups with the combination of continuous and discrete variables (De'ath and Fabricius, 2000).

The partitioning process can be expressed mathematically by the following terms (Washington et al., 1997).

$$D_a = \sum_{l=1}^{L} (Y_{la} - \mu_a)^2 \qquad \text{(Equation 6-1)}$$

where

D_a = total deviance at node a

Y_{la} = l th observation on dependent variable Y in node a

μ_a = mean of L observations in node a

D_a in equation 6-1 is the sum of squared error (SEE) summed over all observations L in node a. More specifically, the squared error term at node a is calculated by the difference between the l th observation of the dependent variable Y and the mean μ of all observations L in node a. As a next step, a split of the observations at a node a occurs on a particular value of an the independent variable X_1 that split into two branches node b and node c. Each node contains M and N observations, portions of the original L observations ($M + N = L$). The D_b and D_c are the residual mean deviances of two branch nodes. Then, the deviation reduction function of equation 6-2 evaluates deviances over all possible independent variable Xs. Thus, the deviance reduction in node a is the greatest when the deviances at node b and node c are the smallest.

$$\Delta_{\text{(all X)}} = D_a - (D_b + D_c) \qquad \text{(Equation 6-2)}$$

$$D_b = \sum_{m=1}^{M} (Y_{mb} - \mu_b)^2 \qquad \text{(Equation 6-3)}$$

$$D_c = \sum_{n=1}^{N} (Y_{nc} - \mu_c)^2 \qquad \text{(Equation 6-4)}$$

where

D_b = total deviation in node b

D_c = total deviation in node c

$\Delta_{(\text{all X})}$ = the total deviance reduction function evaluated over the domain of all Xs

Y_{mb} = mth observation on dependent variable Y in node b

Y_{nc} = nth observation on dependent variable Y in node c

μ_b = mean of L observations in node b

μ_c = mean of L observations in node c

The partitioning process for maximizing deviance is continued at each node until one of the following is met: 1) the node has met minimum population criteria based on statistical sampling theory, or 2) a minimum deviance criterion at a node is met. Previous research has pointed out that the CART analysis techniques have significant advantages compared to the traditional ordinary least-square (OSL) regression model or the logistic regression model as follows (De'ath and Fabricius, 2000; Hallmark, 1999; Lewis, 2000). First, the CART techniques have an intuitive representation so that is simple to interpret. Second, CART analyzer does not need to specify independent variables in advance. Third, the CART is a non-parametric procedure, which does not

need to specify a functional form. So, this is suit for an exploratory knowledge discovery. Fourth, the accuracy of a decision tree is comparable to other models. And finally, CART also allows the exploration of potential interaction effects by tracing variables through branches.

The CART analysis follows the general procedure as below (Lewis, 2000):

1. CART tree building: find the best possible variable to split the data into two sub-data
2. Stopping tree building: continued portioning process until criteria is met
3. Tree pruning: optimal pruning scheme to reduce the complexity and deviation of overall tree structure
4. Optimal tree selection: calculate re-substitution and cross-validation relative error to determine best pruning level

Here, the CART analysis confronts fundamental questions “how to decide the splitting points (tree growing)” and “how to control the size of the tree (tree pruning).” Conceptually, even though the best split to minimize the overall tree deviation is to continue unit further split is impossible, a maximum the tree generally creates overfit problem (Lewis, 2000). Figure 6-1 shows the relationship between tree complexity (the number of terminal nodes considered in tree structure) and the cross-validation error.

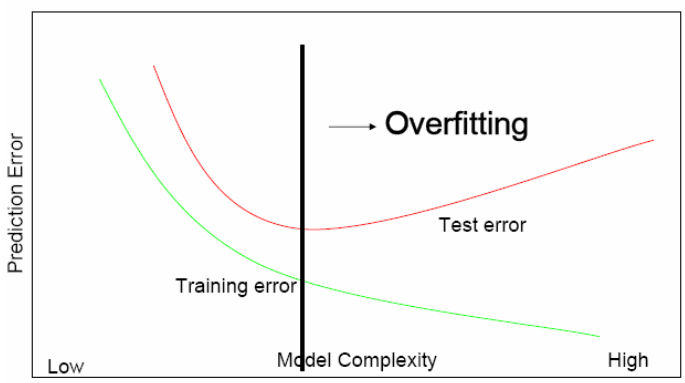


FIGURE 6-1 Training Error and Test Error in CART Analysis (Tsui, 2007)

When the best tree structure applied to an independent subset (test data), the test error with an independent subset (the sum of squared differences of the observations and predictions) does not always decrease. Test data can be used to prune the maximal tree down to an "optimal" tree. Pruning is the process of removing terminal nodes to improve the performance of the decision tree generated using independent data. To find an optimal tree structure, many criteria for stopping the splitting tree process are applied in practical applications. One of the popular approaches is the cross-validation that applies the tree developed from one set of data (usually 90% of original dataset) to another independent data set (10% of original dataset) to evaluate the quality of the tree (StatSoft, Inc., 2007).

6.2 Variable Specifications

After statistical analyses as described in the Chapter Five, this study utilizes the CART analysis in order to determine significant variables and the combination of variables for explaining the difference between VDS and GPS speed. First, the CART

analyses create the “best” tree structure for the response variable (speed difference) both the northbound and the southbound dataset. For each run of the CART analysis, this study identifies the most influential ranges of the predictor variables.

6.2.1 Data Preparation

After the data reduction process as described in the Chapter four, a total of 78,552 individual STN-based trips were generated from the 178 drivers traveled the 37 stations in the GA 400 corridor as shown in Table 6-1.

TABLE 6-1 Data used in the Analysis

	Number of STNs	Number of Trips	Number of Drivers
Southbound	38	41,471	171
Northbound	36	37,081	156
All	74	78,552	178

The model predictor variables used in the CART analysis have four main categories; traffic conditions, roadway characteristics, environmental characteristics, and driver/vehicle characteristics as shown in Table 6-2. All the values of each variable are categorized and converted into discrete variables except the speed limit and the number of lanes. A total amount of 14 variables with 47 choice sets was prepared for the CART analysis.

TABLE 6-2 Factors Hypothesized to Affect Speed and Speed Difference

Type	Factors	Choice	Descriptions
Traffic Conditions	Level of service (LOS)	6	1: A, 2: B, 3: C, 4: D, 5: E, 6: F
	Percentage Truck Traffic	5	1: 0 to 3%, 2: 3 to 5%, 3: 5 to 7%, 4: 7 to 10%, 5: more than 10%
Roadway characteristics	Number of lanes	3	2, 3, and 4
	Speed limit	2	55 and 65 mph
	Freeway sub-type	3	1: basic, 2: on-ramp, 3: off-ramp
Environmental characteristics	Precipitation	2	1: fine day, 2: inclement day
	Daylight	2	1: daylight, 2: darkness
	Time of day	4	1: 6am-10am, 2: 10am-3pm, 3: 3pm-7pm, 4: 7pm -6am
	Weekday/weekend	2	1: weekday, 2: weekend
	Direction	2	1: northbound, 2: southbound
Driver/ Vehicle Characteristics	Driver age	6	1: under 25, 2: 25 to 34, 3: 35 to 44, 4: 45 to 54, 5: 55 to 65, 6: 66 +
	Gender	2	1: male, 2: female
	Vehicle type	4	1: auto, 2: van, 3: SUV, 4: pick-up truck
	Vehicle Age	3	1: less than 5, 2: 5 to 9, 3: 10+
Total		46	

Before apply the CART analysis technique, dataset are divided into two categories by using the factor “direction” in order to avoid the confliction between two factors; the time of day and the direction. For example, while drivers on the southbound corridor experience the most congested traffic during AM period, drivers on the northbound corridor experience the most congested traffic during PM period. In other words, two traffic conditions in the northbound and the southbound have opposite characteristics during the AM and PM periods as shown in Figure 6-2. Thus, when the factors (time of day and direction) are used in the one CART analysis, the results may underestimate or overestimate the effect of two factors.

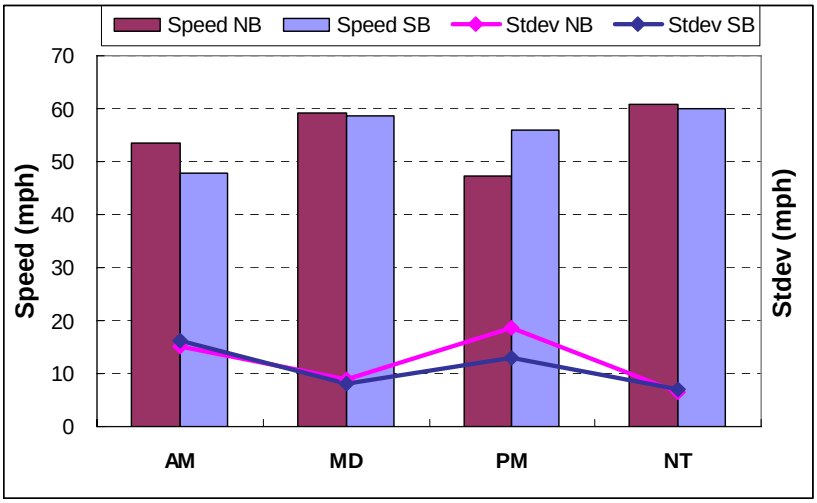


FIGURE 6-2 Mean Speed and Standard Deviation by Direction and Time of Day

The purposes of the CART analysis in this study are to identify variables that significantly contribute to the reduction of the deviation and to determine the magnitude of contribution from significant variables. Theoretically, all possible combinations of 46 choices from 14 factors are 46! (5.5026e+057). Thus, a tree could be constructed

that had one data point at each terminal node. To obtain reasonable results from the analysis, this study set the maximum number of node as 100 and the minimum number of samples as 50 for terminal nodes and as 100 for parent nodes.

6.3 Results of the CART Analysis

After running the CART analysis, the following six main figures are presented for both the northbound and the southbound dataset: 1) best tree structures, 2) cross-validation relative error with addition of nodes, 3) pruned regression tree with detailed information, 4) summary of top 10 splits of best tree structure, 5) number of sample and driver size, and 6) variable importance index.

6.3.1 The Difference Between GPS and VDS Speed in Southbound

The initial regression tree for the southbound dataset was generated using all 13 variables, which consist of 819 nodes. To simplify the regression tree, the prune function that sets 50 for the minimum number of samples at node and 100 for the maximum number of nodes was deployed. After the pruning process, the best tree structure for the southbound dataset had 54 nodes, and total deviation reduced to 0.872.

Figure 6-3 shows the regression tree structure consisting of the top 25 nodes for the southbound dataset. The first split in the regression tree explains the maximum reduction of deviation with subsequent splits explaining consistently less variation. The speed limit was the paramount variable to generate the maximum reduction of the deviation. Driver age was the second significant variable in the dataset and the gender

of drivers and the vehicle type ranked as third and fourth variables to reduce the deviation of the reduction of deviation. From the regression tree with top 25 terminal nodes, most variables had contribution to the difference between GPS and VDS speed. However four variables; number of lane, truck percentage, weather, and weekday/weekend do not show any contribution in regression tree structure for the top 25 terminal nodes.

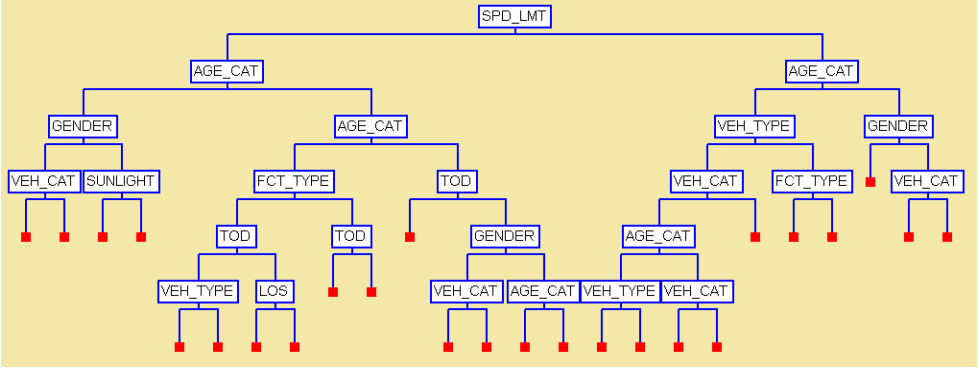


FIGURE 6-3 Regression Trees of the Top 25 Nodes for the Southbound Subset

Figure 6-4 shows the amount of deviance explained by the number of terminal nodes. The best tree structure with 54 nodes had the lowest cross-validation relative error 0.872 and the regression tree with top 25 terminal nodes had 0.882.

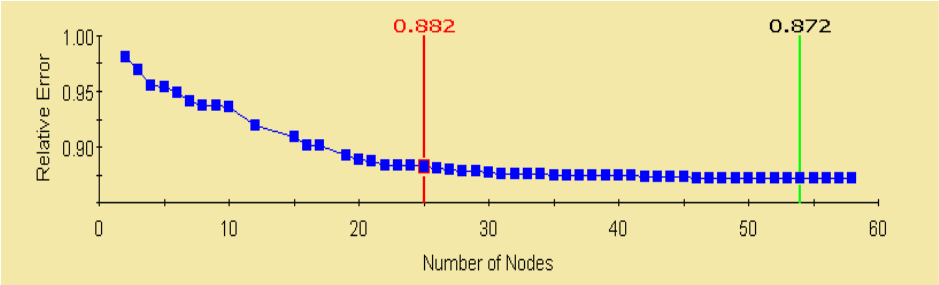


FIGURE 6-4 Cross-validation Relative Error with the Addition of Nodes for the Southbound Subset

Figure 6-5 shows the regression tree structure of top 25 terminal nodes with the detailed information of each node. Parent nodes and 25 terminal nodes have four types of information; node number, variable name, average of speed difference, and number of samples. The speed limit is the first explanatory variable from the best tree structure to split into two homogeneous groups at node 1, which means that the partitioning dataset into the 65 mph speed limit group and the 55 mph speed limit group maximizes the variation reduction from the dataset. More specifically, the mean of the speed difference between GPS and VDS data for original dataset is 9.0 mph. When the dataset is divided into two groups by the speed limit, the mean of the speed difference for 65 mph group is 8.4 mph, while the mean of the speed difference for 55 mph group is 10.9 mph.

The effect of the speed limit in the CART analysis is very similar to the analysis result presented in the Chapter Five. Although the segment of the 55-mph speed limit has lower speed limit than that of the 65-mph speed limit, the mean of the speed difference is higher than the segment of the 65-mph speed limit. The main reason may be that the segment of 65-mph speed limit experiences no geometric changes on the roadways or significant conflicts among vehicles. However, the segment of 55-mph speed limit experiences lane-increases at the two locations from 2 to 3 and 3 to 4 and many conflicts at six on/off-ramps. Hence, this is probably not only speed limit effect itself but interaction with other factors. The speed limit may be correlated with other physical features that did not considered in this analysis. The level of congestion and

freeway sub-type may be the main factors to make higher speed difference on 55-mph speed limit segments. Table 6-3 and 6-4 are summary of the splits for the top 25 nodes from the best tree structure for the southbound dataset.

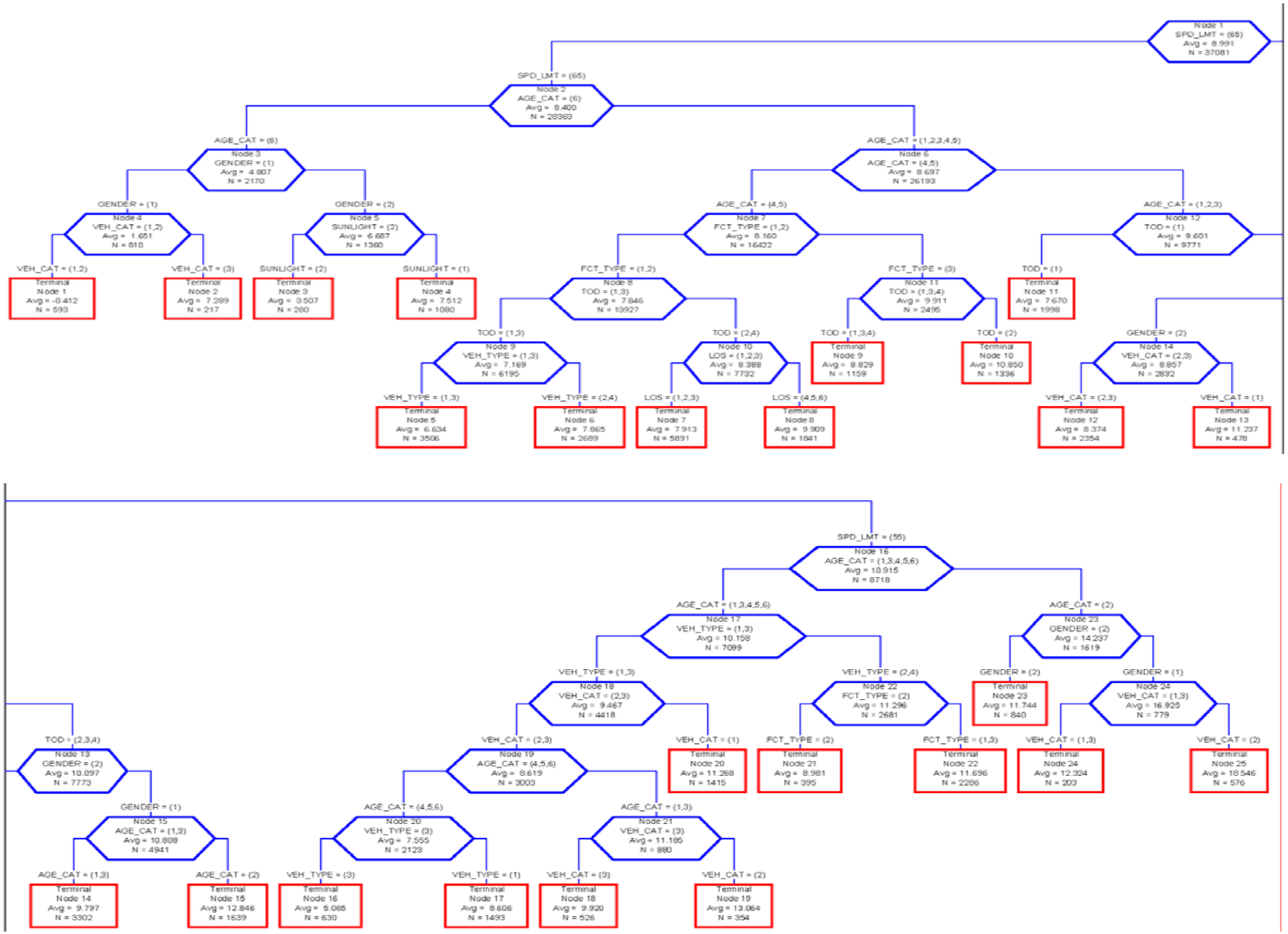


FIGURE 6-5 Trimmed Regression Tree for Speed Difference for Southbound Subset

TABLE 6-3 Summary of Splits from the Best Tree Structure for the Southbound Subset

Rank	Node	Upper Subset Condition Current split	Parent Mean	Sub-node A		Sub-node B	
				Range	Mean	Range	Mean
1	1	Speed limit	9.0	65 mph	8.4	55 mph	10.9
2	2	65mph Driver age	8.4	> 65	4.8	< 65	8.7
3	16	55 mph Driver age	10.9	< 25 or > 35	10.2	25 to 34	14.1
4	6	65mph Driver age	8.7	45 to 64	8.2	< 44	9.6
5	3	65mph / Age > 65 Gender	4.8	Male	1.7	Female	6.7
6	23	55mph / Age 25 to 34 Gender	14.2	Female	11.7	Male	16.9
7	4	65mph / Age > 65 / Female Vehicle year	1.6	< 10	0.4	> 10	7.3
8	7	65mph / Age 45 to 64 Freeway sub-type	8.6	Basic & On-ramp	7.8	Off-ramp	9.9
9	12	65mph / Age < 45 Time of day	9.6	AM	7.7	MD,PM, & NT	10.1
10	13	65mph / Age < 45 / MD,PM,& NT Gender	10.1	Female	8.9	Male	10.8
11	15	65mph / MD,PM,& NT / Male Driver age	10.8	Age < 25 & 35 to 44	9.8	25 to 34	12.8
12	17	55mph / Age < 25 or > 35 Vehicle type	10.2	Auto and SUV	9.5	Van and Truck	11.3
13	18	55mph / Age < 25 & > 35 / Van & SUV Vehicle year	9.5	> 5	8.6	0 to 4	11.3
14	19	55mph / Age < 25 & > 35 / Van & SUV / V_year > 5 Driver age	8.6	> 45	7.6	< 25 & 35 to 44	11.2
15	24	55mph / Age 25 to 34 / Male Vehicle year	16.9	< 5 & >= 10	12.3	5 to 9	18.6
16	20	55mph / Age < 25 & > 35 / Van & SUV / V_year > 5 / Vehicle type	7.6	SUV	5.1	Auto	8.6
17	8	65mph / Age 45 to 64 / Basic & On-ramp Time of day	7.8	AM & PM	7.2	MD & NT	8.4
18	10	65mph / Age 45 to 64 / Basic & On-ramp / MD & NT Level of service	8.4	A to C	7.9	D to F	9.9
19	5	65mph / Age > 65 / Female Sunlight	6.7	Nighttime	3.5	Daytime	7.5
20	14	65mph / Age < 45 / MD,PM,& NT / Female Vehicle year	8.9	> 4	8.4	0 to 4	11.2
21	11	65mph / Age 45 to 64 / Off-ramp Time of day	9.9	AM, PM, & NT	8.8	MD	10.9
22	21	55mph / Van & SUV / V_year > 5 / Age < 25 & 35 to 44 Vehicle year	11.2	>= 10	9.9	5 to 9	13.1
23	9	65mph / Age 45 to 64 / Basic & On-ramp / AM & PM Vehicle type	7.2	Auto & SUV	6.6	Van & truck	7.9

The second to fourth splits are related to the driver age. The second split to maximize the reduction of the deviation from the regression tree structure occurred at the node 2 (65-mph speed limit subset) with the variable driver age. In the 65-mph speed limit segment, drivers older than 65 years old have the speed difference of 4.8 mph, while drivers younger than 65 years old have 8.7 mph speed difference. As similar to the second split, third split occurred at the node 16 (55-mph speed limit segment) with the driver age. The drivers younger than 25 years old or older than 35 years old have 10.2 mph speed difference, while the drivers between 25 to 34 years old have 14.1 mph speed difference. Thus, the drivers older than 65 years old in 65-mph speed limit segment are the least aggressive driver group, but 25 to 34 years old drivers at 55-mph speed limit segment are the most aggressive driver group in the southbound dataset.

Although the mean of the speed difference for the most of the splits for the top 25 nodes appear to be reasonable, several nodes show findings counter to intuitive expectation. For example, in the case of the drivers older than 65 in 65-mph speed limit segment at the node 4, male drivers have 1.7 mph speed difference, while female drivers have 6.7 mph speed difference, which is opposite to finding previous research that male drivers are more commonly associated with a speeding behavior (Jun, 2006; Ko, 2006; Ogle, 2005). Thus, in order to determine whether sample data of each node are representative the population, this study examines the number of drivers for each data subset as shown in Table 6-4.

TABLE 6-4 Samples and Drivers of Top 25 nodes for Southbound Subset

Rank	Node	Sub-node A		Sub-node B	
		Samples	Drivers	Samples	Drivers
1	1	28,363	156	8,718	139
2	2	2,170	24	26,193	132
3	16	7,099	125	1,619	24
4	6	16,422	137	9,771	62
5	3	810	6	1,380	18
6	23	840	8	779	6
7	4	593	5	217	2
8	7	13,927	132	2,495	26
9	12	1,998	16	7,773	61
10	13	2,832	22	4,941	39
11	15	3,302	38	1,639	23
12	17	4,418	35	2,681	26
13	18	3,003	16	1,415	61
14	19	2,123	13	880	6
15	24	203	2	576	3
16	20	630	9	1,493	19
17	8	6,195	62	7,732	57
18	10	5,891	61	9,909	53
19	5	280	4	1,080	17
20	14	2,354	28	478	3
21	11	395	2	2,286	19
22	21	526	3	354	4
23	9	3,506	44	2689	23

The means of speed difference at 10 nodes (node 3, 23, 4, 19, 24, 20, 5, 14, 11, and 21) were calculated by the data from a small number of drivers. Especially, sample data of the node 4 (rank 7) were obtained from six drivers and two drivers. Thus, even though the sunlight effect of the female drivers older than 65 years old in 65-mph speed limit segment at node 5 (rank 19) has the 4 mph speed difference between the daytime,

and the nighttime groups and even though which appears to be reasonable, the variable daylight hardly considered as a significant factor given the small number of drivers. Gender effects of drivers older than 65 years old in the 65-mph speed limit segment at node 3 and drivers from 25 to 34 years old in the 55-mph speed limit segment at the node 19 (rank 14) are also generated from the data of small number of drivers. Finally, the speed difference results from 10 nodes having insufficient number of drivers are also regarded as problematic.

Figure 6-6 shows the relative importance among the 13 variables used in the CART analysis for southbound dataset. In the regression tree of top 25 terminal nodes, the driver age had paramount contribution to the reduction of the deviation compared to rest of the variables.

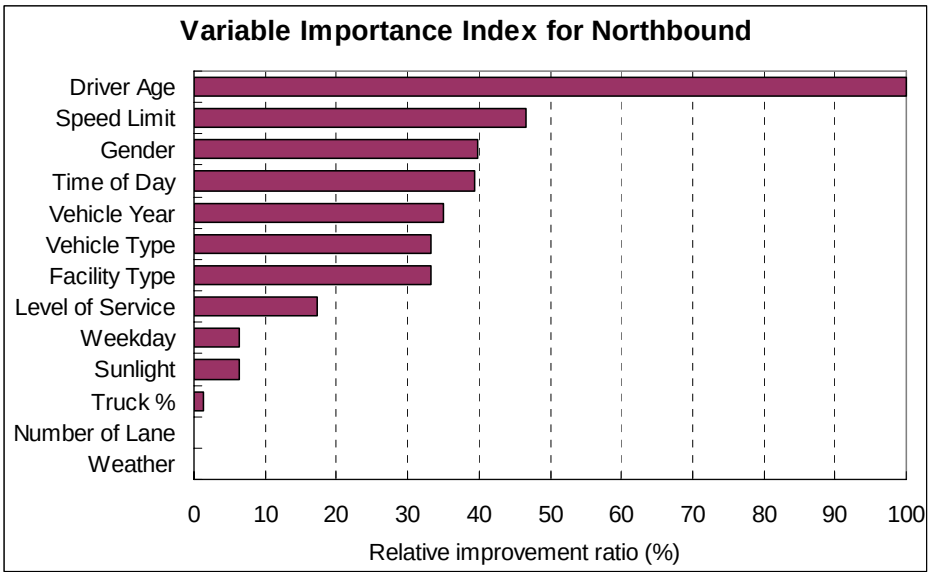


FIGURE 6-6 Variable Importance index for the Southbound Subset

In addition, speed limit, the gender of driver, time of day, vehicle year, vehicle type, and freeway sub-type also had relatively high contribution to the reduction of the deviation. However, the contribution of the second influential variable (speed limit) for the reduction of the deviation is about 57% of most influential variable (driver age), and the contributions of other variables are less than 50% of the variable driver age. However, the five variables (weekday, sunlight, truck percentage, number of lane, and weather) have less than 10% of contribution of the variable driver age. Thus, the driver age is the most significant variable to reduce the deviation of the southbound dataset.

6.3.2 The Difference Between GPS and VDS Speed in the Northbound

As same as the data subset for the southbound, the initial regression tree of the dataset for the northbound was generated by using all 13 variables, which consist of 520 terminal nodes. The prune function was set to 50 for the minimum number of samples for terminal nodes and 100 for maximum number for parent nodes. After pruning process, the best tree structure of the dataset for the northbound had terminal 56 nodes, and the total deviation reduced to 0.864.

Figure 6-7 illustrates regression tree structure consisted of the top 25 terminal nodes for the northbound dataset. The driver age was selected as the paramount variable to maximize reduction of the deviation. The number of lanes and gender ranked as second and third significant variables to reduce the deviation of the difference between GPS and VDS speed. From the regression tree with the top 25 terminal nodes, most variables had contribution to the reduction of deviation. However five variables;

truck percentage, weather, daylight, weekday/weekend, and vehicle year do not show a contribution in the regression tree structure for the top 25 terminal nodes.

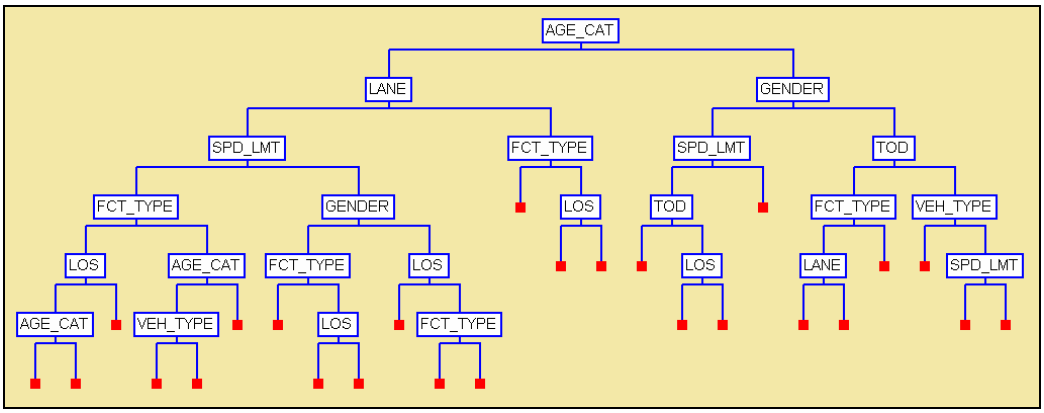


FIGURE 6-7 Regression Trees of the TOP 25 Nodes for the Northbound Subset

Figure 6-8 shows the amount of the deviance explained by the number of terminal nodes. The best tree structure with 56 nodes had the lowest cross-validation relative error 0.864 and the regression tree with the top 25 terminal nodes had 0.879.

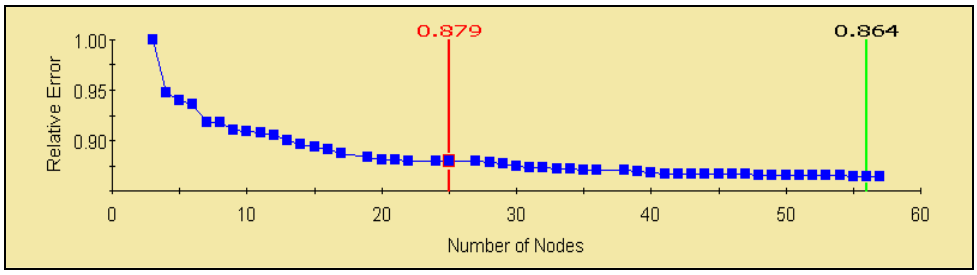


FIGURE 6-8 Cross-Validation Relative Error with the Addition of Nodes for the Northbound Subset

The regression tree structure of the top 25 terminal nodes and parents' nodes is shown in Figure 6-9 with node number, variable name, average speed difference, and number of samples. In addition, Table 6-6 and Table 6-7 summarize the splits for the top 25 nodes from the best tree structure for the northbound dataset. The driver age was the most significant variable from the best tree structure to split into two homogeneous groups at the node 1. More specifically, partitioning dataset into two groups at a driver age of 35 years old maximizes the reduction of variation from the northbound dataset. For example, the mean of the speed difference between GPS and VDS data for the original dataset is 8.6 mph. When the dataset is divided into two driver age groups, the mean of the speed differences for the driver age group older than 35 years old is 8.1 mph, while the mean of speed difference for the driver age group younger than 35 years old is 10.8 mph. Previous research has demonstrated that young drivers not only spend more time speeding, but also speeding at greater extents above the speed limit than other age groups (Ogle, 2005; Ko, 2006). In other words, younger drivers tend to drive faster than average drivers as the result of the node 1.

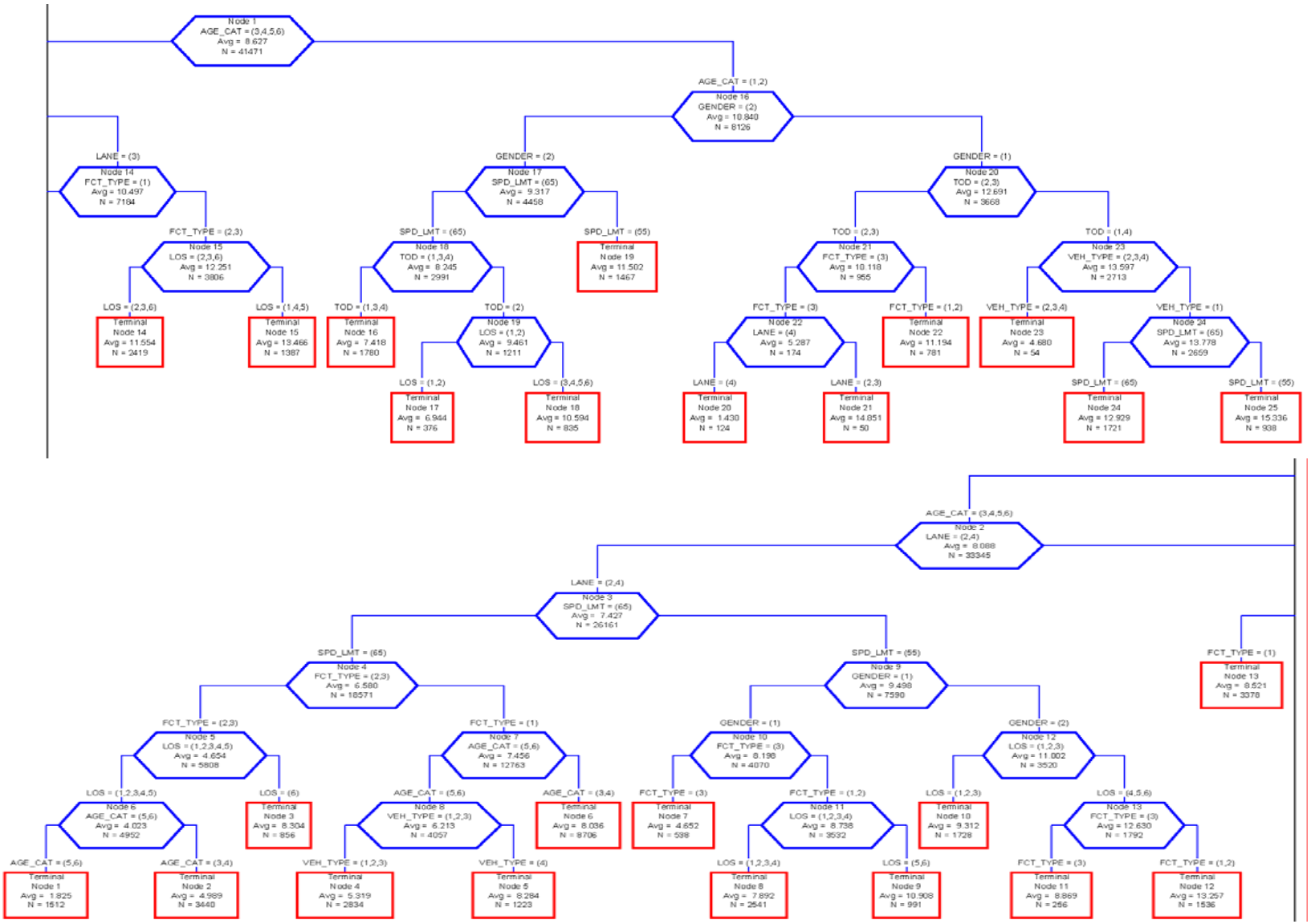


FIGURE 6-9 Trimmed Regression Tree for Speed Difference for the Northbound Subset

TABLE 6-5 Summary of Splits from the Best Tree Structure for the Northbound Subset

Rank	Node	Upper Subset Condition Current split	Parent Mean	Sub-node A		Sub-node B	
				Range	Mean	Range	Mean
1	1	Driver age	8.6	> 35	8.1	< 35	10.8
2	2	Age > 35 Number of Lane	8.1	Lane2 & 4	7.4	Lane 3	10.5
3	3	Age > 35 / Lane 4 Speed Limit	7.4	65 mph	6.6	55 mph	9.5
4	4	Age > 35/ Lane 4 / 65mph Freeway sub-type	6.6	On-ramp & off-ramp	4.7	Basic	7.5
5	14	Age > 35 / Lane 3 Freeway sub-type	10.5	Basic	8.5	On-ramp & off-ramp	12.3
6	16	Age < 35 Gender	10.8	Female	9.3	Male	12.7
7	9	Age > 35 / Lane 4 / 55 mph Gender	9.5	Male	8.2	Female	11.0
8	5	Age> 35 / Lane 4 / 65mph / On/off-ramp Level of service	4.7	A to E	4.0	F	8.3
9	6	Age> 35 / Lane 4 / 65mph / On/off-ramp/ LOS A to E Driver age	4.0	> 55	1.8	35 to 54	5.0
10	17	Age < 35 / Female Speed Limit	9.3	65 mph	8.2	55 mph	11.5
11	12	Age > 35 / Lane 4 / 55 mph / Female Level of service	11.0	A to C	9.3	D to F	12.6
12	7	Age > 35 / Lane 4 / 65mph / Basic Age > 35	7.5	> 55	6.2	35 to 54	8.0
13	20	Age < 35 / Male Time of Day	12.7	MD & PM	10.1	AM & NT	13.6
14	10	Age > 35 / Lane 4 / 55 mph / Male Freeway sub-type	8.2	Off-ramp	4.7	Basic & on-ramp	8.7
15	8	Age < 35/ Male / MD & PM Freeway sub-type	10.1	Orr-ramp	5.3	Basic & on-ramp	11.2
16	11	Age > 35 / Lane 4 / 55 mph / Male / Basic & on-ramp Level of service	8.7	A to D	7.9	E & F	10.9
17	21	Age < 35 / Male / MD & PM Freeway sub-type	10.1	Off-ramp	5.3	Basic & On-ramp	11.2
18	22	Age< 35 / Male / MD & PM / Off-ramp Number of lane	5.3	4-lane-	1.4	2-lane & 3-lane	14.8
19	23	Age <35 / Male / MD & PM Vehicle Type	13.7	Van, SUV, & Truck	4.7	Auto	13.8
20	13	Age > 35 / Lane 4 / 55 mph / Female / LOS D to F Freeway sub-type	12.6	Off-ramp	8.9	Basic & on-ramp	13.3
21	24	Age <35 / Male / MD & PM / Auto Speed limit	13.8	65-mhp	12.9	55-mph	15.3
22	18	Age < 35 / Female / 65 mph Time of day	8.3	AM, PM, & NT	6.9	MD	10.6
23	19	Age < 35 / Female / 65 mph / MD Level of service	9.5	A & B	6.9	C to F	10.6
24	15	Age over 35 / Lane 3 / On/off-ramp Level of service	12.3	B, C & F	11.6	A, D, & E	13.5

The second split to the maximize reduction of the deviation occurred by the number of lanes at the node 2. In the driver age group older than 35 years old, drivers on 2-lane and 4-lane segments have the speed difference of 7.4 mph, while drivers on 3-lane segment have 10.5 mph speed difference. An interaction between several factors may explain this situation. As discussed in the Chapter Five, as traffic condition worsen (e.g., LOS changes from D to F), the difference between GPS and VDS speed increases. The 3-lane segment on the northbound of the study corridor has six on/off-ramps and experienced heavy conflicts between vehicles especially PM peak period compared to other segments. Thus, more congested traffic may cause higher speed difference at 3-lane segment than 2-lane and 4-lane segment.

The third split occurred by variable speed limit at node 3 (rank 3) with the driver age group older than 35 years old. This situation is related to the findings at node 2 because 3-lane segment at node 2 has the speed limit of 55 mph while other segments have the speed limit of 65mph. The finding that the segment of 55 mph speed limit has higher speed difference than that of 65 mph speed limit on northbound was already found at the CART analysis with southbound dataset.

Fourth and fifth split show the interaction between the number of lanes and the freeway sub-types. While fourth split occurred at lane 4 segment under 65 mph speed limit, fifth split occurred at the lane 3 segment under 65 mph speed limit. The speed difference of basic segment is higher than that of on/off-ramp segment in 4-lane segment,

but the speed difference of basic segment is lower than that of on/off segment in 3-lane segment. During the PM period, heavy commuting trips from workplace located in Atlanta downtown area to home flow into the 3-lane segment through three off-ramps. As similar with the node 2, traffic conflicts between vehicles may cause greater speed difference at the six on/off-ramps than basic segment. Thus, the results of split 2 through split 5 are closely related each other. The next split occurred by the variable gender at the node 16 (rank 6) with the driver age group older than 35 years old. Female drivers older than 35 years old have 9.3 mph speed difference, while male drivers older than 35 year old have 12.7 mph speed difference, which appears to be quite reasonable.

The mean of the speed difference for the rest of the splits for top 25 nodes also appear to be reasonable except for several nodes. To examine whether those mean speed differences are representative, this study examines the number of drivers for each data subset. Table 6-7 shows the number of sample size and the number of drivers for each node. The means of speed difference at 10 nodes (node 9, 17, 12, 21, 22, 23, 13, 18, 19, and 15) were calculated by the data of the small number of drivers. Especially, sample data of the node 22 (rank 18) were obtained from four and two drivers. Thus, the speed difference results from 10 nods having insufficient number of drivers are regarded as problematic results.

TABLE 6-6 Sample and Drivers of Top 25 nodes for the Northbound Subset

Rank	Node	Sub-node A		Sub-node B	
		Samples	Drivers	Samples	Drivers
1	1	33,345	143	8,126	28
2	2	26,161	143	7,148	117
3	3	18,571	125	7,590	113
4	4	5,808	87	12,763	130
5	14	3,378	47	3,806	53
6	16	4,556	38	3,668	25
7	9	4,070	28	3,520	21
8	5	4,952	62	856	12
9	6	1,512	35	3,440	41
10	17	955	13	2,713	12
11	12	1,728	23	1,792	27
12	7	4,057	86	8,706	121
13	20	955	19	2,713	31
14	10	538	11	3,532	72
15	8	2,834	38	1,223	20
16	11	2,541	31	991	16
17	21	174	13	781	13
18	22	124	4	50	2
19	23	54	2	2,659	34
20	13	256	5	1,536	12
21	24	1,721	31	938	17
22	18	1,780	24	1,211	19
23	19	376	6	835	8
24	15	2,419	29	1,387	21

Figure 6-10 shows the relative importance among 13 variables used in the CART analysis for the northbound dataset. In the regression tree of the top 25 terminal nodes, the freeway sub-type has paramount contribution to the reduction of the deviation compared to the rest of the variables. In addition, speed limit, driver age, and number of lanes have high contribution to the reduction of the deviation, which are greater than 80% contribution of the freeway sub-type. Gender and level of service have greater than 50% contribution of the most influential variable (freeway sub-type). However, four

variables (sunlight, weekend, truck percentage, and weather) have less than 10% of contribution of the variable driver age. Thus, freeway sub-type, speed limit, driver age, and number of lanes are major significant variables to reduce the deviation of the northbound subset.

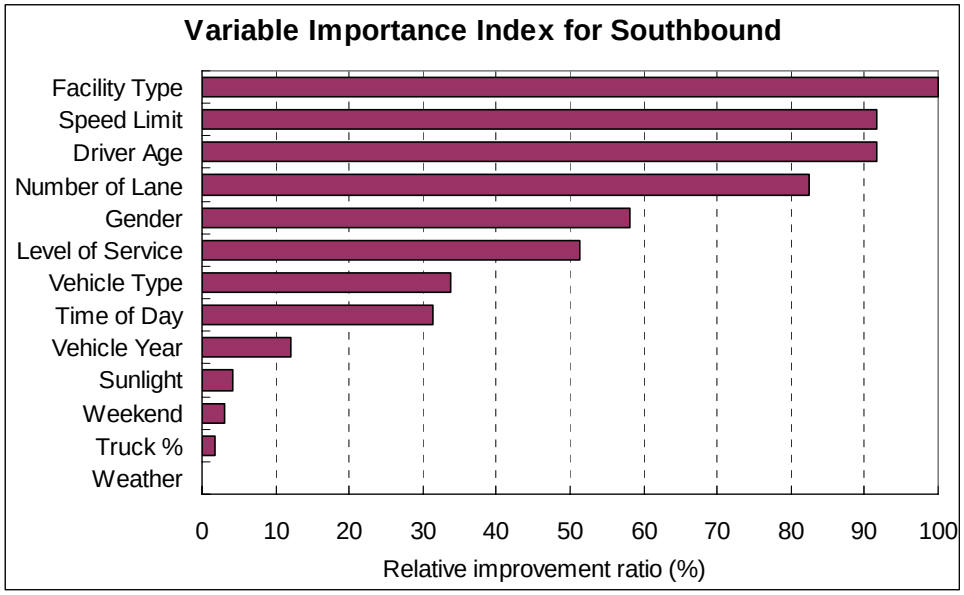


FIGURE 6-10 Variable Importance Index for the Northbound Subset

In southbound the dataset, driver age was the most dominant variable to reduction of the deviation compared to other variables. However, in the northbound dataset, four variables have similar contribution for the reduction of the deviation.

6.4 Summary

This chapter utilized the CART analysis to identify significant variables and the combination of variables that may explain the difference between VDS and GPS speed. After the generating best tree structure for both the southbound and the northbound datasets, the most influential ranges of the variables were identified with the pruned tree structure of the top 25 terminal nodes. Then, significant variables were selected by applying a criterion 30 as the minimum number of drivers for each subset. The CART analyses show that the driver age was the most relevant variable in explaining variation for the southbound dataset and freeway sub-type, speed limit, driver age, and number of lane were the most influential variables for the northbound dataset. Table 6-3 and Table 6-6 show the detailed information regarding speed difference between sub-groups. The combination of several variables had significant contribution in the reduction of the deviation for both the northbound and the southbound dataset. Even though this study generates the relationship between speed difference and various factors, the results of the CART analysis should be considered with the driver sample size to acquire statistically significant results. Expanded sampling with larger number of drivers would enrich this study results.

CHAPTER SEVEN

CONCLUSION AND FURTHER RESEARCH

7.1 Summary of the Findings

The goal of this study was to evaluate VDS speed data used to estimate freeway travel speeds using GPS-equipped vehicle data. Analyses focused on the differences noted between GPS and VDS data as a function of traffic congestion levels, roadway characteristics, environmental conditions, and driver/vehicle characteristics. Preliminary analysis shows that the mean of GPS speeds is higher than VDS speeds, and the standard deviation of the GPS speed is usually equal to or lower than the VDS deviations (for most time periods and locations). This research also examined the potential use of global positioning system (GPS) data as an alternative means to provide data for non-working VDS stations.

The difference between GPS and VDS speeds is affected by various factors such as the roadway and traffic conditions, environmental characteristics, and drivers/vehicles characteristics. All of those factors are also closely related to the VDS data accuracy. The following is a list of the findings of this study.

- As LOS worsens, the speed difference increase and VDS data accuracy may worsen of potential GPS sampling bias may increase.

- As truck percentage in traffic increases from less than 3% to 3-5%, the speed difference increase and VDS data accuracy may worsen.
- As number of lanes increase, the speed difference increase and VDS data accuracy may worsen or GPS sampling bias increase.
- 55mph speed limit segment had greater speed difference than 65 mph speed limit segment due to the congested traffic condition and heavy conflicts between vehicles, which means VDS data accuracy in 55 mph speed limit segment may be worse than 65 mph speed limit segment.
- Basic segments had greater speed difference than on/off-ramps, which means that worse VDS accuracy may be found in basic segment than on/off-ramp segment.
- The speed difference of day time is greater than that of night time, which is opposed to the expectation. However traffic congested condition generally occurred during daytime rather than at night time. All vehicles at night hardly experienced much congestion compared to the vehicles during the day time.
- The AM period had the biggest speed difference from four time periods, and the NT period had the smallest speed difference, which means that VDS accuracy is closely related with the traffic congestion.

This study utilized classification and regression tree (CART) analysis in order to determine significant variables and the combination of variables for explaining the difference between VDS and GPS speed. CART analysis results found that that the driver age was the most relevant variable in explaining variation for the southbound

freeway dataset and freeway sub-type, speed limit, driver age, and number of lane were the most influential variables for the northbound freeway dataset. The combination of several variables had significant contribution in the reduction of the deviation for both the northbound and the southbound dataset. Even though this study generates the relationship between speed difference and various factors, the results of the CART analysis should be considered with the driver sample size to acquire statistically significant results. Expanded sampling with larger number of drivers would enrich this study results.

7.2 Limitations and Further Research

To obtain more reliable results, VDS camera characteristics such as VDS camera installation and calibration information should be included for the investigation of the difference between GPS and VDS speed. The difference between GPS and VDS speed may be affected by the VDS camera calibration and installation because the VDS speed accuracy depends on those two key elements. By doing so, best parameters for the camera calibration and maintenance can be determined by the analysis with GPS data.

Another possibly application of GPS data is the integration GPS/GIS techniques with the VDS real-time traffic information for the real-time route and congestion monitoring system. Phase III of the Commute Atlanta Study will be monitoring vehicle location, current congestion level based on current speed from VDS real-time information, and communicating the congestion price to the driver in real-time.

REFERENCES

Brilon, W. and Ponzlet, M., 1996. Variability of Speed-Flow Relationships on German Autobahns. Transportation Research Record(1555): 91-98.

De'ath, G. and Fabricius, K.E., 2000. Classification and Regression Trees: a Powerful yet Simple Technique for Ecological Data Analysis. ECOLOGY -NEW YORK-: VOL 81; PART 11, p. 3178-3192.

Faghri, A. and Hamad, K., 2002. Travel Time, Speed, and Delay Analysis Using an Integrated GIS/GPS System. Canadian Journal of Civil Engineering, 29(2): 325-328.

Garcia, C., Huebschman, R., Abraham, D.M. and Bullock, D.M., 2006. Using GPS to Measure the Impact of Construction Activities on Rural Interstates. Journal of Construction Engineering and Management: Vol. 132, No. 5, pp. 508-515.

Georgia State DOT, 2007. Real-time Traffic Map.

Grant, C., Gillis, B. and Guensler, R., 1999. Collection of Vehicle Activity Data by Video Detection for Use in Transportation Planning. ITS Journal: Vol. 5, Issue 4, pp. 47.

Greaves, S. and Somers, A., 2003. Insights on Driver Behaviour: What Can Global Positioning System (GPS) Data Tell Us? Proceedings - Conference of the Australian Road Research Board. ARRB Transport Research Ltd., Vermont South, VIC. 3133, Australia, Cairns, QLD, Australia, pp. 2993-3007.

Hallmark, S.L., 1999. Analysis and Prediction of Individual Vehicle Activity for Microscopic Traffic Modeling, Ph.D. Thesis, The Georgia Institute of Technology, Atlanta, GA.

Hallmark, S.L., Knapp, K.K. and Grant, C.D., 2004. Evaluating Speed Differences Between Cars, Light-duty Trucks, and Vans for Emissions Modeling. Journal of Transportation Engineering: Vol. 130, Issue 6, pp. 814-817.

Hoogendoorn, S.P., 2005. Vehicle-type and Lane-specific Free Speed Distributions on Motorways: A Novel Estimation Approach Using Censored Observations. Transportation Research Record(1934): 148-156.

Hori, T., 1997. Traffic Camera System Development, Proceedings of SPIE-The International Society for Optical Engineering. Proc. SPIE - Int. Soc. Opt. Eng. (USA). Vol. 3028, 1997, pp. 81-90, San Jose, CA, USA.

Hostovsky, C., Wakefield, S. and Hall, F.L., 2004. Freeway Users' Perceptions of Quality of Service: Comparison of Three Groups. Transportation Research Record(1883): 150-157.

Jun, j., 2006. Potential Crash Exposure Measures Based on GPS-Observed Driving Behavior Activity Metrics, Ph.D. Thesis, The Georgia Institute of Technology, Atlanta, GA.

Kanellaidis, G., 1995. Factors Affecting Drivers' Choice of Speed on Roadway Curves. Journal of Safety Research: Vol. 26, Issue 1, pp. 49-56.

Klein, L.A., 1993. Traffic Parameter Measurement Technology Evaluation. IEEE, Piscataway, NJ, USA, Ottawa, Ontario, Canada, pp. 529-533.

Ko, j., 2006. Measurement of Freeway Traffic Flow Quality Using GPS-Equipped Vehicles, Ph.D. Thesis, The Georgia Institute of Technology, Atlanta, GA.

Kyte, M., Khatib, Z., Shannon, P. and Kitchener, F., 2001. Effect of Weather on Free-flow Speed. Transportation Research Record(1776): 60-68.

Lewis, R.J., 2000. An Introduction to Classification and Regression Tree (CART) Analysis, Presented to 2000 Annual Meeting of the Society for Academic Emergency Medicine, San Francisco, California, USA.

Liang, W.L., Kyte, M., Kitchener, F. and Shannon, P., 1998. Effect of Environmental Factors on Driver Speed: A Case Study. Transportation Research Record(1635): 155-161.

Martin, P. and Feng, Y., 2003. Detector Technology Evaluation, Mountain Plains Consortium and Utah Department of Transportation

Martin, P. and Stevanovic, A., 2004. Evaluation of UDOT's Video Detection Systems - System's Performance in Various Test Conditions, Utah Department of Transportation and Department of Civil and Environmental Engineering.

Middleton, D. and Parker, R., 2004. Initial Evaluation of Selected Detectors to Replace Inductive Loops on Freeways, Texas Transportation Institute.

Mimbela, L.E.Y. and Klein, L.A., 2000. Summary of Vehicle Detection and Surveillance Technologies Used in Intelligent Transportation Systems, New Mexico University and Federal Highway Administration.

Minnesota DOT and SRF Consulting Group, 2002. NIT Phase II: Evaluation of Non-intrusive Technologies for Traffic Detection.

Ogle, J., 2005. Quantitative Assessment of Driver Speeding Behavior Using Instrumented Vehicles, Ph.D. Thesis, The Georgia Institute of Technology, Atlanta, GA.

Ogle, J., Guensler, R., Bachman, W., Koutsak, J. and Wolf, J., 2002. Accuracy of GPS for Determining Driver Performance Parameters. Transportation Research Record: No. 1818, pp. 12-24.

Oregon DOT, 2005. Travel Time Messaging on Dynamic Message Signs-Portland.

Recarte, M.A. and Nunes, L., 2002. Mental Load and Loss of Control Over Speed in Real Driving: Towards a Theory of Attentional Speed Control. Transportation Research Part F: Traffic Psychology and Behaviour, 5(2): 111-122.

Rhodes, A., Bullock, D.M., Sturdevant, J., Clark, Z. and Candey Jr, D.G., 2005. Evaluation of the Accuracy of Stop Bar Video Vehicle Detection at Signalized Intersections. Transportation Research Record(1925): 134-145.

Schneider, W. and Mrakotsky, E., 2005. Mobile Phones as a Basis for Traffic State Information. IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC. Institute of Electrical and Electronics Engineers Inc., Piscataway, NJ 08855-1331, United States, Vienna, Austria, pp. 782-784.

Texas Transportation Institute and Cambridge Systematics Inc., 2005. Traffic Congestion and Reliability: Trends and Advanced Strategies for Congestion Mitigation, Federal Highway Administration.

Tian, Z., 2006. Critical Elements of Video Detection System Applications at Signalized Intersections. Proceedings of the Conference on Traffic and Transportation Studies, ICTTS. American Society of Civil Engineers, Reston, VA 20191-4400, United States, Xi'an, China, pp. 576-583.

TMC Pooled-Fund Study, 2007. TMC Performance Monitoring: Evaluation and Reporting Handbook.

TRB, N.R.C., 2000. Highway Capacity Manual.

Tsui, K., 2007. Lecture Notes: Data Mining, "Data Mining & Modeling".

URS Corporation and GeoStats Inc., 2003. NaviGAtor Data Archive Improvements.

Wang, Z., 2004. Using floating cars to measure travel time delay. Transportation Research Record: No. 1870, 84-93.

Washington, S., 2000. Iteratively Specified Tree-based Regression: Theory and Trip Generation Example. Journal of Transportation Engineering, 126(6): 482-491.

Washington State DOT, 2007. How Accurate are WSDOT's Calculated Travel Times?

Wolf, J., Guensler, R., Washington, S. and Bachman, W., 1998. High-emitting Vehicle Characterization Using Regression Tree Analysis. Transportation Research Record(1641): 58-65.

Wolf, J., Schonfelder, S., Samaga, U., Oliveira, M. and Axhausen, K.W., 2004. Eighty Weeks of Global Positioning System Traces: Approaches to Enriching Trip Information. Transportation Research Record(1870): 46-54.

Ygnace, J.L. and Drane, C., 2001. Cellular Telecommunication and Transportation Convergence: A Case Study of a Research Conducted in California and in France on Cellular Positioning Techniques and Transportation Issues. IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC, Oakland, CA, pp. 16-22.

Zito, R. and Taylor, M.A.P., 1994. Use of GPS in Travel-time Surveys. Traffic Engineering & Control: Vol. 35, Issue 12, pp. 685-690.

APPENDEX A

The Kolmogorov-Smirnov statistics (KS statistics) represent the maximum difference between the two distributions. If KS statistics are greater than the p-value, the test rejects the hypothesis that the two distributions are the not different at the 5% confidence level (Mathworks, 2007).

Table A-1 K-S Statistics and P-values for the Pair-wise Comparisons of Speed Difference Distributions by Truck Percentage

	0-1	1-3	3-5	5-7	7-10	10+
0-1	-	0.048 (0.037)	0.037 (0.019)	0.034 (0.016)	0.051 (0.003)	0.070 (0.003)
1-3		-	0.028 (0.009)	0.029 (0.019)	0.035 (0.022)	0.035 (0.002)
3-5			-	0.013 (0.016)	0.018 (0.002)	0.039 (0.000)
5-7				-	0.022 (0.011)	0.073 (0.001)
7-10					-	0.046 (0.000)
10+						-

Table A-2 K-S Statistics and P-values for the Pair-wise Comparisons of Speed Difference Distributions by number of Lanes

	2 Lanes	3 Lanes	4 Lanes
2 Lanes	-	0.077 (0.000)	0.089 (0.000)
3 Lanes		-	0.037 (0.000)
4 Lanes			-

Table A-3 K-S Statistics and P-values for the Pair-wise Comparisons of Speed Difference Distributions by Speed Limit

	55 MPh	65 MPH
55 MPH	-	0.208 (0.000)
65 MPH		-

Table A-4 K-S Statistics and P-values for the Pair-wise Comparisons of Speed Difference Distributions by Freeway Sub-type

	Basic Segment	On-ramp	Off-ramp
Basic	-	0.105 (0.000)	0.213 (0.000)
On-ramp		-	0.220 (0.000)
Off-ramp			-

Table A-5 K-S Statistics and P-values for the Pair-wise Comparisons of Speed Difference Distributions by Weather

	non-inclement	Inclement day
non-inclement	-	0.039 (0.000)
Inclement day		-

Table A-6 K-S Statistics and P-values for the Pair-wise Comparisons of Speed Difference Distributions by Daylight

	Daylight	Night	Twilight
Day	-	0.073 (0.019)	0.068 0.000)
Night		-	0.049 (0.000)
Twilight			-

Table A-7 K-S Statistics and P-values for the Pair-wise Comparisons of Speed Difference Distributions by Time of Day

	AM	MD	PM	NT
AM	-	0.105 (0.000)	0.090 (0.000)	0.077 (0.000)
MD		-	0.021 (0.007)	0.048 (0.000)
PM			-	0.036 (0.005)
NT				-

Table A-8 K-S Statistics and P-values for the Pair-wise Comparisons of Speed Difference Distributions by Weekday and Weekend

	Weekday	Weekend
Weekday	-	0.013 (0.030)
Weekend		-